Personality and Alignment Processes in Dialogue: Towards a Lexically-Based Unified Model

Carsten Brockmann

Doctor of Philosophy

Institute for Communicating and Collaborative Systems

School of Informatics

University of Edinburgh

2009

Abstract

This thesis explores approaches to modelling individual differences in language use. The differences under consideration fall into two broad categories: Variation of the *personality* projected through language, and modelling of language *alignment* behaviour between dialogue partners. In a way, these two aspects oppose each other – language related to varying personalities should be recognisably different, while aligning speakers agree on common language during a dialogue.

The central hypothesis is that such variation can be captured and produced with restricted computational means. Results from research on personality psychology and psycholinguistics are transformed into a series of lexically-based Affective Language Production Models (ALPMs) which are parameterisable for personality and alignment. The models are then explored by varying the parameters and observing the language they generate.

ALPM-1 and ALPM-2 re-generate dialogues from existing utterances which are ranked and filtered according to manually selected linguistic and psycholinguistic features that were found to be related to personality. ALPM-3 is based on true overgeneration of paraphrases from semantic representations using the OPENCCG framework for Combinatory Categorial Grammar (CCG), in combination with corpus-based ranking and filtering by way of n-gram language models. Personality effects are achieved through language models built from the language of speakers of known personality. In ALPM-4, alignment is captured via a cache language model that remembers the previous utterance and thus influences the choice of the next. This model provides a unified treatment of personality and alignment processes in dialogue.

In order to evaluate the ALPMs, dialogues between computer characters were generated and presented to human judges who were asked to assess the characters' personality. In further internal simulations, cache language models were used to reproduce results of psycholinguistic priming studies. The experiments showed that the models are capable of producing natural language dialogue which exhibits human-like personality and alignment effects.

Acknowledgements

First and foremost, I would like to thank Jon Oberlander, who was an inspiring, reliable and generous supervisor. His feedback and encouragement throughout the course of this thesis were invaluable. I am also grateful to my second supervisor Johanna D. Moore for discussions that helped to shape my work, and to my examiners Colin Matheson and Paul Piwek for their useful comments and for a stimulating viva.

I am indebted to Jon Oberlander, Amy Isard and Michael White, who worked with me on the Critical Agent Dialogue project (CrAg). The thesis emerged from this project; I gratefully acknowledge funding by Scottish Enterprise as part of the Edinburgh-Stanford Link. Special thanks go to Amy Isard for compiling, transcribing and annotating the CrAg Corpus and for implementing the version of the CrAg system that uses the OPENNLP CCG Library (OPENCCG); and to Michael White for developing OPENCCG and for collaboration that led to the cache language modelling approach.

Interesting discussions with, and feedback from, other staff and fellow students in the School of Informatics are also gratefully acknowledged. In particular, I am indebted to Alastair Gill and Scott Nowson for results and resources of their research on language and personality, which were integrated into the models developed in this thesis. Thanks also go to Mary Ellen Foster and, again, Michael White for resources from the FLIGHTS system and the COMIC project, and to Oliver Lemon for comments on an earlier version of this dissertation.

On a personal note, I would like to thank my friends and family, especially my parents, Gudrun and Rolf Brockmann, for all their love and support over the years.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Carsten Brockmann*)

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Objectives 1

1.3 Limitations 2

1.4 Thesis Plan 2

2 Background and Related Work **5**

2.1 Summary 5

2.2 Personality 5

2.2.1 Personality Models 5

2.2.2 Personality and Language Behaviour 7

2.3 Alignment 9

2.3.1 Priming 9

2.3.2 The Interactive Alignment Model of Dialogue Processing . . 10

2.3.3 Corpus-Based Priming and Alignment Studies 11

2.3.4 Sociolinguistic Research on Accommodation 11

2.4 Personality and Alignment in Computational Linguistics 12

2.4.1 Personality Recognition from Text 12

2.4.2 Generating Personality-Rich Language 14

2.4.3 Stylistically Controlled Generation 17

2.4.4 Human–Computer Interaction 18

2.4.5 Generation of Dialogues between Computer Characters 18

3 Recognising Personality from Text **21**

3.1 Summary 21

3.2 Introduction 22

3.2.1 Hypotheses 22

3.3 Methods . . . 23
3.3.1 CrAg Corpus . . . 23
3.3.2 Feature Combination . . . 28
3.3.3 Re-Generating Dialogue . . . 30
3.3.4 Materials . . . 34
3.4 Results . . . 40
3.5 Discussion . . . 43

4 Generating Text That Projects Personality 45
4.1 Summary . . . 45
4.2 Introduction . . . 46
4.2.1 Hypothesis . . . 46
4.3 Methods . . . 47
4.3.1 OpenCCG Grammar . . . 47
4.3.2 N-Gram Language Models . . . 48
4.3.3 Personality Models . . . 49
4.3.4 Base Language Model . . . 50
4.3.5 Character Specification and Dialogue Generation . . . 50
4.3.6 Materials . . . 54
4.4 Results . . . 65
4.5 Discussion . . . 67

5 Modelling Alignment with Cache Language Models 73
5.1 Summary . . . 73
5.2 Introduction . . . 74
5.2.1 Hypothesis . . . 75
5.3 Methods . . . 75
5.3.1 Cache Language Models . . . 75
5.3.2 Interpolating Cache and Default Language Models . . . 76
5.4 Experiment 1: Repetition of Nouns and Adjectives . . . 78
5.4.1 Introduction . . . 78
5.4.2 Methods . . . 80
5.4.3 Results . . . 84
5.4.4 Discussion . . . 89
5.5 Experiment 2: Semantic Relatedness . . . 93
5.5.1 Introduction . . . 93

5.5.2 Methods . . . 94
5.5.3 Results . . . 95
5.5.4 Discussion . . . 98
5.6 Experiment 3: Prepositional Object/Double Object Priming . . . 100
5.6.1 Introduction . . . 100
5.6.2 Methods . . . 103
5.6.3 Results . . . 104
5.6.4 Discussion . . . 106
5.7 Discussion . . . 108

6 A Unified Model of Personality and Alignment 111
6.1 Summary . . . 111
6.2 Introduction . . . 111
6.2.1 Hypotheses . . . 112
6.3 Methods . . . 112
6.3.1 Natural Language Generation Guided by Cache Language Models . . . 112
6.3.2 Materials . . . 113
6.4 Results . . . 117
6.5 Discussion . . . 124

7 Conclusion 129
7.1 Summary . . . 129
7.2 Contributions . . . 130
7.3 Future Work . . . 132

A Experimental Instructions 133
A.1 Experiment on the Perception of Communication Styles . . . 133
A.2 Experiment on the Perception of Communication Styles II . . . 136

B Experimental Results 141
B.1 Modelling Alignment with Cache Language Models . . . 141
B.1.1 Experiment 1: Repetition of Nouns and Adjectives . . . 141
B.1.2 Experiment 2: Semantic Relatedness . . . 150
B.1.3 Experiment 3: Prepositional Object/Double Object Priming . . . 155

Bibliography 161

List of Figures

3.1 Partial additive multiattribute value function (AMVF) for extravert language. 29
3.2 Boxplot for the personality scores computed for the CrAg Corpus utterances available for re-generation. 35
3.3 Percentage of native English speakers' choices in agreement with expectations, by personality dimension. 41
3.4 Percentage of non-native English speakers' choices in agreement with expectations, by personality dimension. 42
3.5 Percentage of all participants' choices in agreement with expectations, by personality dimension. 42

4.1 Simple utterance specification. 52
4.2 Fragment of the logical form generated from the utterance specification in Figure 4.1. 53
4.3 Mean Spearman's rank correlation ρ at varying personality model weight settings: Five-factor model and extraversion (E) dimension. 55
4.4 Mean Spearman's rank correlation ρ at varying personality model weight settings: Five-factor model and neuroticism (N) dimension. 56
4.5 Mean Spearman's rank correlation ρ at varying personality model weight settings: CrAg Corpus individual *gq*'s language model and E dimension. 56
4.6 Mean Spearman's rank correlation ρ at varying personality model weight settings: CrAg Corpus individual *sd*'s language model and N dimension. 57
4.7 Percentage of native English speakers' choices in agreement with expectations, by personality dimension. 68
4.8 Percentage of non-native English speakers' choices in agreement with expectations, by personality dimension. 68

4.9 Percentage of all participants' choices in agreement with expectations, by personality dimension. 69
4.10 Native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went. 69
4.11 Non-native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went. 70
4.12 All participants' ratings of how well the characters got on with each other and how smoothly the conversation went. 70

5.1 OPENCCG logical form of the parse of phrase (5.10)(a). 81
5.2 OPENCCG logical form of the parse of phrase (5.10)(b). 82
5.3 Mean geometric mean scores of pre-nominal targets following one of eight different primes. 85
5.4 Mean probabilities of pre-nominal targets following one of eight different primes. 86
5.5 Mean geometric mean scores of relative clause targets following one of eight different primes. 88
5.6 Mean probabilities of relative clause targets following one of eight different primes. 89
5.7 Mean geometric mean scores of pre-nominal targets following one of six different primes. 96
5.8 Mean geometric mean scores of relative clause targets following one of six different primes. 97
5.9 Mean geometric mean scores of prepositional object targets following one of four different primes. 105
5.10 Mean geometric mean scores of double object targets following one of four different primes. 106

6.1 Percentage of native English speakers' choices in agreement with expectations, by personality dimension, with alignment. 120
6.2 Percentage of non-native English speakers' choices in agreement with expectations, by personality dimension, with alignment. 120
6.3 Percentage of all participants' choices in agreement with expectations, by personality dimension, with alignment. 121
6.4 Percentage of native English speakers' choices in agreement with expectations, by personality dimension, without and with alignment. . . 121

6.5 Percentage of non-native English speakers' choices in agreement with expectations, by personality dimension, without and with alignment. . 122
6.6 Percentage of all participants' choices in agreement with expectations, by personality dimension, without and with alignment. 122
6.7 Native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went, with alignment. . 123
6.8 Non-native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went, with alignment. 123
6.9 All participants' ratings of how well the characters got on with each other and how smoothly the conversation went, with alignment. 124
6.10 Native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went, without and with alignment. 126
6.11 Non-native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went, without and with alignment. 127
6.12 All participants' ratings of how well the characters got on with each other and how smoothly the conversation went, without and with alignment. 128

List of Tables

2.1 Traits associated with Eysenck's three personality dimensions, adapted from Matthews et al. (2003, p. 22). 7
2.2 Trait facets associated with Costa and McCrae's five personality dimensions, adapted from Matthews et al. (2003, p. 24). 8
2.3 Summary of language cues for extraversion with corresponding PERSONAGE generation parameters, reproduced from Mairesse and Walker (2007, p. 498). 16

3.1 Topics assigned to the utterances in the CrAg Corpus. 24
3.2 Sentiment polarities assigned to the utterances in the CrAg Corpus. . . 24
3.3 CrAg Corpus overall utterance counts by topic. 25
3.4 CrAg Corpus context-independent utterance counts by topic. 26
3.5 Features characteristic of high extravert language, with corresponding component value functions. 31
3.6 Features characteristic of high neurotic language, with corresponding component value functions. 32
3.7 CrAg Corpus utterances available for re-generation. 34
3.8 Personality parameter settings for the four experimental conditions. . 36
3.9 Exemplars for the personality dimensions displayed during the web experiment. 37
3.10 Numbers of expected versus not expected choices for the web experiment. 41

4.1 Topics available in the CrAg OpenCCG grammar, with nouns used to realise them. 47
4.2 NEO-PI-R scores for the persons in the CrAg Corpus. 51
4.3 Topics and polarities for the pretest to determine the personality model weight that maximises variety. 54

4.4 Spearman's rank correlation ρ for the test utterances generated with high versus low E and N settings, for both five-factor and individual personality language models. 58
4.5 Personality parameter settings for the four experimental conditions. . 60
4.6 Exemplars for the personality dimensions displayed during the web experiment, taken from Goldberg (1992, p. 34 f.). 62
4.7 Numbers of expected versus not expected choices for the web experiment. 66

5.1 Example semantic classes assigned to lexical items. 77
5.2 Proportions of prepositional object (PO) and double object (DO) target responses following PO or DO prime completions. Results of Experiment 1 from Pickering and Branigan (1998, p. 639, Table 1). 101

6.1 Personality and alignment parameter settings for the four experimental conditions. 113
6.2 Numbers of expected versus not expected choices for the web experiment, with results of binomial tests to determine whether the distributions differ from chance. 118
6.3 Comparison of getting on and smoothness ratings without and with alignment, for native English speakers, non-native English speakers and all speakers. 125

B.1 Mean geometric mean (GM) scores (with standard deviations) of pre-nominal targets following pre-nominal (P-N) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). 142
B.2 Mean geometric mean (GM) scores (with standard deviations) of pre-nominal targets following relative clause (RC) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). 143
B.3 Mean geometric mean (GM) scores (with standard deviations) of relative clause targets following pre-nominal (P-N) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). 144

B.4 Mean geometric mean (GM) scores (with standard deviations) of relative clause targets following relative clause (RC) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). 145
B.5 Mean probabilities (with standard deviations) of pre-nominal targets following pre-nominal (P-N) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). 146
B.6 Mean probabilities (with standard deviations) of pre-nominal targets following relative clause (RC) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). 147
B.7 Mean probabilities (with standard deviations) of relative clause targets following pre-nominal (P-N) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). 148
B.8 Mean probabilities (with standard deviations) of relative clause targets following relative clause (RC) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). 149
B.9 Mean geometric mean (GM) scores (with standard deviations) of pre-nominal targets following pre-nominal (P-N) or relative clause (RC) primes with same noun (SN), semantically related noun (SRN) or semantically unrelated noun (SUN). 151
B.10 Mean geometric mean (GM) scores (with standard deviations) of relative clause targets following pre-nominal (P-N) or relative clause (RC) primes with same noun (SN), semantically related noun (SRN) or semantically unrelated noun (SUN). 152
B.11 Mean probabilities (with standard deviations) of pre-nominal targets following pre-nominal (P-N) or relative clause (RC) primes with same noun (SN), semantically related noun (SRN) or semantically unrelated noun (SUN). 153
B.12 Mean probabilities (with standard deviations) of relative clause targets following pre-nominal (P-N) or relative clause (RC) primes with same noun (SN), semantically related noun (SRN) or semantically unrelated noun (SUN). 154
B.13 Mean geometric mean (GM) scores (with standard deviations) of prepositional object targets following prepositional object (PO) or double object (DO) primes with same verb (SV) or different verb (DV). . . . 156

B.14 Mean geometric mean (GM) scores (with standard deviations) of double object targets following prepositional object (PO) or double object (DO) primes with same verb (SV) or different verb (DV). 157
B.15 Mean probabilities (with standard deviations) of prepositional object targets following prepositional object (PO) or double object (DO) primes with same verb (SV) or different verb (DV). 158
B.16 Mean probabilities (with standard deviations) of double object targets following prepositional object (PO) or double object (DO) primes with same verb (SV) or different verb (DV). 159

List of Acronyms

A	agreeableness
ALPM	Affective Language Production Model
ALPM-1	Affective Language Production Model version 1
ALPM-2	Affective Language Production Model version 2
ALPM-3	Affective Language Production Model version 3
ALPM-4	Affective Language Production Model version 4
AMVF	additive multiattribute value function
ANOVA	analysis of variance
C	conscientiousness
CAT	Communication Accommodation Theory
CCG	Combinatory Categorial Grammar
CLM	cache language model
COMIC	COnversational Multimodal Interactions with Computers project
CrAg	Critical Agent Dialogue project
CRAG 1	Critical Agent Dialogue system version 1
CRAG 2	Critical Agent Dialogue system version 2
DO	double object
DV	different verb

E	extraversion
EAR	Electronically Activated Recorder
ECA	embodied conversational agent
EPQ-R	Eysenck Personality Questionnaire-Revised
FLIGHTS	Fancy Linguistically Informed Generation of Highly Tailored Speech system
GM	geometric mean
IAM	Interactive Alignment Model
IC	Intolerable Cruelty
LIWC	Linguistic Inquiry and Word Count
LM	language model
LSM	linguistic style matching
LXG	League of Extraordinary Gentlemen
MR	Mystic River
MRC	Medical Research Council
MRCPD	MRC Psycholinguistic Database
MXPOST	Maximum entropy model for Part-Of-Speech Tagging
N	neuroticism
NECA	Net Environment for Embodied Emotional Conversational Agents
NEO-PI-R	NEO-Personality Inventory-Revised
NLG	natural language generation
NLP	natural language processing
O	openness
OAA	Open Agent Architecture

OPENCCG	OPENNLP CCG Library
OPENNLP	open source projects related to natural language processing
P	psychoticism
PAULINE	Planning And Uttering Language In Natural Environments
PERSONAGE	Personality Generator
PERSONAGE-PE	Personality Generator-Parameter Estimation
PN	pre-nominal
PO	prepositional object
POS	part-of-speech
RC	relative clause
ROC	Rank Order Centroid
RRL	Rich Representation Language
SD	standard deviation
SN	same noun
SPARKY	Sentence Planning with Rhetorical Knowledge
SRILM	SRI Language Modeling Toolkit
SRN	semantically related noun
SUN	semantically unrelated noun
SV	same verb
SVM	support vector machine
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

Chapter 1

Introduction

1.1 Motivation

It is known that a speaker's personality influences their language use. For example, extraverts, who like to be the center of attention, tend to produce long sentences and to talk about events in a positive way, whereas introverts have a tendency to talk about themselves and to express themselves more negatively. These personality effects are supposed to be stable.

On the other hand, in dialogue, it has been found that a speaker aligns their language to their conversational partner. For the duration of the dialogue, speakers form 'conversational pacts' and agree on lexical items used to refer to entities during their discussion. More particularly, when hearing the other person speak, priming effects lead to repetition of syntactic structures; for example, if there is a choice of a double object or a prepositional object construction during a picture description experiment, participants tend to re-use the construction they just heard. In order to achieve a successful dialogue, speakers dynamically change their linguistic and situational representations.

Personality and alignment effects on language production are therefore in tension. This thesis sets out to explore this tension using computational linguistic methods.

1.2 Objectives

Both personality and alignment have been studied in separation in computational linguistic and psycholinguistic research. The objective of this thesis is to work towards a lexically-based *unified* computational model which enables us to generate language,

and in particular dialogue, that emulates human personality characteristics as well as alignment effects.

Throughout the thesis, a series of Affective Language Production Models (ALPMs) will be proposed, and their parameters will be systematically varied to generate textual dialogues between computer characters discussing a movie. The goal is for the linguistic personalities of the characters to be clearly identifiable, for adaptation to be possible, and for the interaction between them to be believable and engaging. The models will be evaluated by human judges in web-based experiments, and with computational simulations of psycholinguistic studies.

This research is of interest to psycholinguists, who do not yet have an implemented computer model of alignment processes; researchers in natural language processing, who are currently seeking more flexible models to underpin more principled natural language generation systems; and researchers in social responses to computing technology, who have shown that simple language manipulations can influence a user's perception of the personality implicit in a computer interface, but who cannot yet experiment with more thorough-going language effects.

1.3 Limitations

In order to limit the scope of the thesis, we intend to investigate the capabilities of *lexically-based* approaches, i.e., models which exploit knowledge associated with lexical items, but do not require information about syntactic structure.

As features relating personality and language use have been identified in previous research, we will use these results to inform our models instead of endeavouring to learn new features from scratch with machine learning approaches.

Furthermore, the generated dialogues will consist only of text. Both personality and alignment effects can be found in single modalities, as well as in multimodal interactions. We will not attempt to generate animated movies of embodied conversational agents with speech output and gesture manipulations.

1.4 Thesis Plan

Chapter 2 provides an overview of related work relevant to the thesis.

In Chapter 3, we propose a model to rank arbitrary unseen utterances by personality, informed by Oberlander and Gill's Affective Language Production Model (ALPM-

1 and ALPM-2) and by the Critical Agent Dialogue project (CrAg) Corpus, a collection of human dialogues in the movie review domain. The model uses a variety of machine-readable linguistic and psycholinguistic resources, which are combined according to features determined in previous research on the relationship of language and personality. For evaluation purposes, dialogues between two computer characters are generated by re-combining utterances from the CrAg Corpus according to the rank that the model assigns to them when personality parameters are altered systematically. These dialogues are then presented to humans in a personality perception web experiment.

Chapter 4 introduces ALPM-3, a significantly more flexible model that generates a set of paraphrases from a semantic representation and ranks these paraphrases using n-gram language models to choose an utterance that best matches personality parameter settings. Again, dialogues between computer characters are generated and presented to human judges in order to determine whether differences are recognisable.

While the previous two chapters exclusively dealt with the modelling of personality effects, Chapter 5 introduces a matching approach based on cache language models (CLMs) that is designed to model priming and alignment effects. Findings from three psycholinguistic experiments on the repetition of nouns and adjectives, on semantic relatedness and on prepositional object/double object priming are replicated in experiments in order to determine how adequately the cache language modelling approach is able to emulate human behaviour.

Chapter 6 unifies the matching approach with the ALPM-3 personality model into ALPM-4, which is capable of generating dialogues between computer characters that vary in both personality and alignment behaviour. Dialogues are presented to human judges once more, and results are compared to those without alignment.

In Chapter 7, we summarise the contributions and outline possible avenues for future work.

Chapter 2

Background and Related Work

2.1 Summary

This chapter serves as a high-level overview of relevant work related to the topics of this thesis. We consider studies in the area of personality psychology in relation to language behaviour, psycholinguistic research on priming and alignment, and sociolinguistic theories of accommodation. We then discuss work in computational linguistics: Machine-learning approaches to the modelling of personality, results on personality and alignment in human–computer interaction, and systems which generate dialogues between computer characters. We motivate how these areas inspired the models developed throughout the thesis and point forward to later chapters where appropriate.

2.2 Personality

2.2.1 Personality Models

In personality psychology, there are two major models which classify personality according to either three or five higher-order dimensions. A detailed discussion of the theories and their position in personality trait research can be found in Matthews et al. (2003); here, we will only briefly summarize the approaches.

The Affective Language Production Models (ALPMs) version 1 and 2, proposed in Chapter 3, are based on hypotheses related to the three-factor model of personality, while ALPM-3 and ALPM-4, introduced in Chapters 4 and 6, make use of corpora annotated with five dimensions.

2.2.1.1 Eysenck's Three-Factor Model

Eysenck's personality model (Eysenck and Eysenck 1975, Eysenck et al. 1985) comprises the three dimensions extraversion (E), neuroticism (N) and psychoticism (P). They are assessed by a self-report questionnaire, the Eysenck Personality Questionnaire-Revised (EPQ-R), a collection of yes/no questions. The questionnaire also contains a scale to determine whether the person is lying.

In our first web-based personality perception experiment (see Section 3.3.4.2), we used paraphrased descriptions from Eysenck and Eysenck (1975) to introduce the dimensions to the study's participants. The technical terms *neuroticism* and *psychoticism* were replaced by *emotional instability* and *tough-mindedness*, respectively, to make them more accessible to laymen and to avoid negative connotations.

Extraversion Typical extraverts are sociable, like parties, have many friends, need to have people to talk to, and do not like reading or studying by themselves. They crave excitement, take chances, often stick their neck out, act on the spur of the moment, and are generally impulsive individuals. They are fond of practical jokes, always have a ready answer, and generally like change; they are carefree, easy-going, optimistic, and like to "laugh and be merry." They prefer to keep moving and doing things, tend to be aggressive and lose their temper quickly; altogether their feelings are not kept under tight control, and they are not always reliable people.

Emotional instability Typical emotionally unstable people are anxious, worrying individuals, moody and frequently depressed. They are likely to sleep badly, and to suffer from various psychosomatic disorders. They are overly emotional, reacting too strongly to all sorts of stimuli, and find it difficult to get back on an even keel after each emotionally arousing experience. Their strong emotional reactions interfere with their proper adjustment, making them react in irrational, sometimes rigid ways. If emotionally unstable individuals have to be described in one word, one might say that they are *worriers*; their main characteristic is a constant preoccupation with things that might go wrong, and a strong emotional reaction of anxiety to these thoughts.

Tough-mindedness Tough-minded individuals may be described as being solitary, not caring for people; they are often troublesome, not fitting in anywhere. They may be cruel and inhumane, lacking in feeling and empathy, and altogether insensi-

Dimension	Traits
Extraversion	Sociable, lively, active, assertive, sensation seeking, carefree, dominant, surgent, venturesome.
Neuroticism	Anxious, depressed, guilt feelings, low self-esteem, tense, irrational, shy, moody, emotional.
Psychoticism	Aggressive, cold, egocentric, impersonal, impulsive, antisocial, unempathetic, creative, tough-minded.

Table 2.1: Traits associated with Eysenck's three personality dimensions, adapted from Matthews et al. (2003, p. 22).

> tive. They are hostile to others, even with their own kith and kin, and aggressive even to loved ones. They have a liking for odd and unusual things, and a disregard for danger; they like to make fools of other people and to upset them.

Table 2.1 lists traits associated with the three dimensions, as found in Matthews et al. (2003, p. 22).

2.2.1.2 The Big Five Model

Costa and McCrae's personality model (Costa and McCrae 1992) suggests the five dimensions extraversion (E), neuroticism (N), agreeableness (A), conscientiousness (C) and openness (O). Due to their widespread acceptance, these are sometimes called the *Big Five*. The dimensions are assessed by the NEO-Personality Inventory-Revised (NEO-PI-R) self-report questionnaire. Table 2.2 lists trait facets associated with the five dimensions, according to Matthews et al. (2003, p. 24).

The three-factor model's psychoticism dimension is replaced by agreeableness, conscientiousness and openness in the Big Five model. For a direct comparison of the three- and five-factor models, see Matthews et al. (2003, pp. 25 ff.).

2.2.2 Personality and Language Behaviour

Scores on the personality dimensions have been found to be related to language behaviour. Dewaele and Furnham (1999) studied the influence of extraversion on linguistic variation. In subsequent work, Heylighen and Dewaele (2002) introduced the

Dimension	Trait Facets
Extraversion	Warmth, gregariousness, assertiveness, activity, excitement seeking, positive emotions.
Neuroticism	Anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability.
Agreeableness	Trust, straightforwardness, altruism, compliance, modesty, tender-mindedness.
Conscientiousness	Competence, order, dutifulness, achievement striving, self-discipline, deliberation.
Openness	Fantasy, aesthetics, feelings, actions, ideas, values.

Table 2.2: Trait facets associated with Costa and McCrae's five personality dimensions, adapted from Matthews et al. (2003, p. 24).

technical notion of *formality* which is associated with preference for, for instance, nouns and adjectives, as opposed to verbs and adverbs; and preference for formality is related to level of extraversion.

Pennebaker and colleagues analysed the linguistic style of texts using their program Linguistic Inquiry and Word Count (LIWC, Pennebaker et al. 2001). Pennebaker and King (1999) reported correlations of LIWC categories with the five-factor personality dimensions.

Gill and Oberlander studied the linguistic projection of personality and derived features characteristic for extraversion, neuroticism and psychoticism (Gill and Oberlander 2002, Gill 2004, Oberlander and Gill 2004, 2006). These form the basis of the model for recognition of personality from text developed in Chapter 3; the list of features we use to recognise extravert or neurotic language can be found in Section 3.3.2.

Nowson (2006) applied Gill's methods to discover individual differences in weblog corpora and also investigated gender effects. Nowson's corpora inform ALPM-3 and ALPM-4 (Chapters 4 and 6).

Personality can not only be projected through, but also perceived from asynchronous textual communication. Gill (2004) and Gill et al. (2006) found that the E dimension was perceived most accurately, and P to a lesser extent, while it was difficult for raters to recognise N. The authors propose that this difficulty stems from characteristics

of the N dimension and from the study's setting, computer-mediated communication at zero-acquaintance.

While these previous studies mainly focused on the *discovery* of features relating personality and language behaviour, one of the contributions of this thesis is to apply this knowledge in order to *generate* recognisable personality variation.

2.3 Alignment

2.3.1 Priming

Above, we established that a speaker's personality is related to individual differences in language production. In contrast to this, however, speakers also tend to repeat their own and each other's linguistic choices, both in monologue and in dialogue. This process is called *priming*. The repetition of syntactic structures is referred to as *structural priming* or *syntactic priming*. Syntactic structures that have been studied in the context of priming include the prepositional object (PO)/double object (DO) alternation (Bock 1986, 1989, Bock and Loebell 1990, Pickering and Branigan 1998, Branigan et al. 2000), noun phrases (Cleland and Pickering 2003) or passives (Bock 1986, Bock and Loebell 1990).

Viewed as an experimental paradigm, priming is flexible and reliable and yields strong results. There are different experimental setups, e.g., picture description experiments, sentence recall or sentence completion. Priming occurs within language production as well as from comprehension to production. Effects can be found in a variety of languages and between languages in bilinguals. Pickering and Ferreira (2008) provide a comprehensive review of research on these and other aspects of structural priming.

Bock (1986) presented experimental participants with priming sentences in PO or DO form (e.g., sentences (2.1)(a) and (b)) or active or passive form (e.g., sentences (2.1)(c) and (d)) and then asked them to describe a target picture (e.g., of a man reading a story to a boy or a church being struck by lightning). Each of the four syntactic forms elicited a higher percentage of replies in the same form than in the alternative form.

(2.1) (a) A rock star sold some cocaine to an undercover agent.

(b) A rock star sold an undercover agent some cocaine.

(c) One of the fans punched the referee.

(d) The referee was punched by one of the fans.

Later, Bock (1989) argued that structural priming cannot be explained only in terms of lexical repetition of closed-class words. Primes were prepositional *to*- or *for*-datives (e.g., sentences (2.2)(a) and (b)) or double object *to*- or *for*-datives (e.g., sentences (2.2)(c) and (d)). Priming of PO or DO forms occurred regardless of which preposition the prime contained.

(2.2) (a) A cheerleader offered a seat to her friend.

(b) A cheerleader saved a seat for her friend.

(c) A cheerleader offered her friend a seat.

(d) A cheerleader saved her friend a seat.

Priming occurs without content-word repetition, but when there is such repetition, the priming effect is boosted. Semantic relatedness of prime and target also increases the priming effect (Cleland and Pickering 2003).

Gill et al. (2004) conducted an experiment using the confederate scripting technique (Branigan et al. 2000) to study the influence of personality on priming effects. They found that a moderate level of neuroticism facilitated priming and a high level of N inhibited it, while extraversion had no influence.

2.3.2 The Interactive Alignment Model of Dialogue Processing

Priming plays a central role in Pickering and Garrod's Interactive Alignment Model of dialogue processing (IAM, Garrod and Pickering 2004, Pickering and Garrod 2004, 2006). According to this theory, each dialogue participant uses a set of internal representations, corresponding to different linguistic levels, i.e., phonological, syntactic and semantic levels. They also have a representation of a situation model that represents a particular state of affairs in the dialogue.

For successful communication, interlocutors align their situation models, and this happens through the largely unconscious, automatic process of *interactive alignment*. There is a parity of representations used in production and comprehension, and there is priming of representations between speakers and listeners. This leads to imitation of words, sounds, grammatical forms and meanings used by the dialogue partner, which

in turn causes alignment of the interlocutors' representations. Alignment at linguistically lower levels leads to more alignment at other levels, and ultimately to alignment of the situation models.

2.3.3 Corpus-Based Priming and Alignment Studies

In addition to psycholinguistic experiments, priming was studied with corpus-based methodologies (e.g., Dubey et al. 2005, Gries 2005, Szmrecsanyi 2005, Reitter 2008).

Dubey et al. (2005) argue that the parallelism preference effect is an instance of a general priming mechanism. The parallelism effect is based on psycholinguistic findings and specifies that speakers process the second conjunct of a coordinate construction faster if it has the same internal syntactic structure as the first conjunct.

Reitter (2008) developed metrics and found evidence for short-term priming and long-term adaptation effects in the Switchboard and HCRC Map Task dialogue corpora. Long-term adaptation predicted communicative success, while short-term priming did not. Reitter also found evidence that priming is sensitive to syntactic structure: There was less priming of pairs of part-of-speech categories that crossed constituent boundaries (*distituents*) than of pairs that occurred within constituents.

Niederhoffer and Pennebaker (2002) found evidence for linguistic style matching (LSM) in dyadic interactions on both the conversation level as well as on a turn-by-turn level.

2.3.4 Sociolinguistic Research on Accommodation

Linguistic adaptation to interlocutors has also been studied from a sociolinguistic perspective. A prominent approach is Communication Accommodation Theory (CAT), which was surveyed by Shepard et al. (2001).

According to CAT, individuals use language to achieve a desired social distance between themselves and interacting partners. They employ a number of strategies to achieve this. In the context of this thesis, we are interested in the *convergence* strategy, which refers to the modification of accents, dialects, idioms or features such as speech rate, pauses and utterance length. Interlocutors converge in order to be like those to whom they are attracted, and also in order to ensure that an interaction flows smoothly. People may also choose to *diverge* in order to emphasise distinctiveness from the other. Convergence is generally rated positively by the individual that is its target, while divergence is often perceived negatively.

The power structure of the relationship between two interactants can affect whether there is upward or downward accommodation. There can be partial, full, or even over-accommodation. While alignment is assumed to be largely automatic, accommodation is usually considered to be a conscious process, at least in part. At times, the speaker's intention may be inconsistent with their behaviour, or the listener may not perceive or misinterpret a speaker's accommodation behaviour. A certain level of accommodation may be expected in interactions; convergence on all dimensions, under- or overaccommodation lead to miscommunication. For example, in endeavouring to be 'down with the kids', an adult can attempt to adopt a teenager's sublanguage, and end up overaccommodating, by using language even more characteristic of the target subgroup than that which the teenager would dare use themselves.

In Chapters 5 and 6, we contribute a parameterisable computational model of alignment, which enables us to replicate human behaviour and explore the effects of varying levels of alignment on the perception of dialogues. The results are then compared to the psycholinguistic, corpus linguistic and sociolinguistic findings reported above.

2.4 Personality and Alignment in Computational Linguistics

2.4.1 Personality Recognition from Text

This thesis is concerned with the modelling of personality and alignment effects in natural language with computational linguistic approaches. In related work, Mairesse et al. used machine learning methods to train statistical models of personality ratings (Mairesse and Walker 2006a,b, Mairesse et al. 2007).

For the study reported in Mairesse and Walker (2006a), the training data were collected by Mehl et al. (2006). Random bits of 96 participants' conversations were recorded while they were wearing an Electronically Activated Recorder (EAR), and were transcribed afterwards. Five to seven independent observers scored the transcriptions on the dimensions of the five factor model of personality (extraversion (E), neuroticism (N), agreeableness (A), conscientiousness (C) and openness (O)); these scores were averaged for use by the system.

For each utterance under consideration, four sets of features were extracted automatically:

- Linguistic Inquiry and Word Count (LIWC, Pennebaker et al. 2001) features. (See Section 3.3.1.3.3 for a description of LIWC.)
- MRC Psycholinguistic Database (MRCPD, Wilson 1988) features. (See Section 3.3.1.3.2 for a description of the MRCPD.)
- Utterance type features (ratio of commands, prompts or back-channels, questions or assertions).
- Prosodic features (e.g., speech rate or pitch or intensity of the voice).

Personality models expressed as rules were trained using RankBoost.

Compared to a baseline model that ranks extracts randomly, a model trained with all features performed significantly better for the E, A, C and O dimensions, while N was the most difficult dimension to model. Using only the LIWC feature set produced similar results. The prosodic features alone performed best at modelling E, the MRCPD features alone were best for N, and the LIWC features alone were best for O; A and C were modelled best by the all feature model. For the best-performing models, the authors list the features that have the most impact on the recognition of personality traits. Previous findings are confirmed and new markers are identified.

Mairesse and Walker (2006b) also examined statistical models for the five factor model of personality and used the same feature sets as Mairesse and Walker (2006a) but modified training data and machine learning algorithms. In addition to the observer personality ratings of the EAR corpus, self-reports on the same data were included, as well as a corpus of essays from psychology students, collected by Pennebaker and King (1999), with self-assessed personality information.

Five regression models of personality dimensions as continuous variables were computed: Linear regression, M5′ regression tree with linear models, M5′ decision tree with regular leaves, REP-Tree decision tree and support vector machines (SVMs). In a second task, six classification models were used to split subjects into high or low groups for each personality dimension: J48 decision tree, nearest neighbour, Naive Bayes, JRip rules set, AdaboostM1 and SMO SVMs.

The regression models were trained and compared to a baseline model returning the mean of all personality scores in the training set. For the corpus of essays, significant but small improvements over the baseline were found for all personality dimensions with both linear regression and M5′ regression tree with linear models, and O was the easiest dimension to model. For the EAR corpus with observer ratings, models

other than linear regression significantly improved over the baseline for the E, N and C dimensions, while for EAR self-reports no models achieved significance.

In the classification case, models improved significantly over a baseline returning the majority class for the E, N and C dimensions with the EAR corpus with observer ratings, and for the E and O dimensions with the EAR self-reports. Training a Naive Bayes classifier on separate feature sets showed that LIWC features performed well for the E and N dimensions, MRCPD features for E and C, prosodic features for E and especially O and speech acts for A. Overall, the authors concluded that perceived personality was easier to model than projected personality and that spoken language was easier to model than written text.

Argamon et al. (2005) attempted to classify authors as High or Low Extravert and High or Low Neurotic, using Pennebaker and King's (1999) data. They reported classification accuracies of around 58% (with a 50% baseline). Oberlander and Nowson (2006) undertook a comparable task, using weblog data. They reported classification accuracies of roughly 75% (E), 84% (N), 81% (A) and 82% (C). The weblog corpus authors' scores for O were not normally distributed, which is why this dimension was excluded from the study.

Our approach to the recognition of personality from text, as detailed in ALPM-1 and ALPM-2 in Chapter 3, does not attempt to learn personality features with machine learning approaches. Instead, it relies on, and is informed by, previous studies.

A further difference to Mairesse et al.'s work is that throughout this thesis, our assumptions about features and our corpus resources are solely based on projected, i.e., self-assessed personality data, which are arguably more accurate than observer ratings. If the only goal is to model extreme personalities with maximum recognisability, then it suffices to use perceived personality data; but if another goal is realism, then projection is worth investigating.

2.4.2 Generating Personality-Rich Language

Sentence Planning with Rhetorical Knowledge (SPARKY, Walker et al. 2003) is a natural language generation (NLG, Reiter and Dale 2000) system that generates spoken language for recommendations or comparisons in the restaurant domain. Its input is a content plan, which consists of a set of assertions of facts about the restaurant under discussion and a specification of rhetorical relations that hold between the facts. A sentence plan generator transforms the content plan into competing sentence plans, which

are represented by sets of automatically generated features that count occurrences of certain structural configurations.

Based on the SPARKY architecture, Mairesse and Walker (2007) developed Personality Generator (PERSONAGE), a rule-based system for generating personality-rich language either directly or by overgeneration and selection. Psychological findings on the extraversion dimension were mapped to the components of the NLG system – content planning, sentence planning and realisation (see Table 2.3).

PERSONAGE was used to generate 240 utterances; 40 with introvert, 40 with extravert and 160 with random parameter settings. Three judges rated each of the utterances on a scale from 1 to 7 for perceived extraversion and naturalness. Interrater agreement, measured as the average Pearson's correlation between the judges' ratings, was 0.57. The average ratings for introvert and extravert sentences were significantly different, which indicated that the manipulation could be recognised.

Mairesse and Walker also trained regression models on the judges' ratings to obtain models that could assign extraversion ratings to utterances. The correlation between the best model's output and personality ratings was higher than the correlations between pairs of judges, from which the authors conclude that the model matches human performance.

From PERSONAGE, Mairesse and Walker (2008) developed Personality Generator-Parameter Estimation (PERSONAGE-PE). Two human judges rated a set of 160 randomly generated utterances on each of the Big Five personality dimensions. Statistical models were then trained on this annotated data to predict 67 generation parameters. For each parameter, the best model was selected via cross-validation. In comparison with PERSONAGE, this approach is no longer rule-based and does not overgenerate and rank utterances. Instead of varying only extraversion, it can be used to produce variation on all five dimensions.

For evaluation, 50 utterances were generated, each of which targeted extreme (high or low) values for two personality dimensions and neutral values for the other three. Human judges were then asked to judge these utterances on the respective extreme dimensions (without being told that they were extreme) and on naturalness. The correlation between the model's target personality scores and the judges' ratings was highest for extraversion (0.45) and lowest for conscientiousness (-0.01), which was the only non-significant correlation. Average ratings for the utterances generated with high or low target values differed significantly for all dimensions except conscientiousness, which means that those individual dimensions could be perceived even with the utter-

NLG modules	Introvert findings	Extravert findings	Parameter	Intro	Extra
Content selection and structure	Single topic	Many topics	Verbosity	low	high
	Strict selection	Think out loud*	Restatements	low	high
			Repetitions	low	low
	Problem talk, dissatisfaction	Pleasure talk, agreement, compliment	Content Polarity	low	high
			Repetitions Polarity	low	high
			Claim Polarity	low	high
			Concessions	avg	avg
			Concessions Polarity	low	high
			Polarisation	low	high
			Positive Content First	low	high
Syntactic template selection	Few self-references	Many self-references	Self-references	low	high
	Elaborated constructions	Simple constructions*	Claim Complexity	high	low
	Many articles	Few articles			
Aggregation operations	Many words per sentence/clause	Few words per sentence/clause	Relative Clauses	high	low
			With Cue Word	high	low
			Conjunction	low	high
	Many unfilled pauses	Few unfilled pauses	Period	high	low
			...		
Pragmatic transformations	Many nouns, adjectives, prepositions (explicit)	Many verbs, adverbs, pronouns (implicit)	Subject Implicitness	low	high
	Many negations	Few negations	Negation Insertion	high	low
	Many tentative words	Few tentative words	Downtoner Hedges:		
			· *sort of*, *somewhat*, *quite*, *rather*, *err*, *I think that*, *it seems that*, *it seems to me that*, *I mean*	high	low
			· *around*	avg	avg
	Formal	Informal	· *kind of*, *like*	low	high
			Acknowledgments:		
			· *yeah*	low	high
			· *right*, *OK*, *I see*, *well*	high	low
	Realism	Exaggeration*	Emphasizer Hedges:		
			· *really*, *basically*, *actually*, *just have*, *just is*, exclamation	low	high
			· *you know*	low	high
	No politeness form	Positive face redressment*	Tag Question Insertion	low	high
	Lower word count	Higher word count	Hedge Variation	low	avg
			Hedge Repetition	low	low
Lexical choice	Rich	Poor	Lexicon Frequency	low	high
	Few positive emotion words	Many positive emotion words	*See polarity parameters.*		
	Many negative emotion words	Few negative emotion words	*See polarity parameters.*		

Table 2.3: Summary of language cues for extraversion with corresponding PERSONAGE generation parameters, reproduced from Mairesse and Walker (2007, p. 498). Asterisks indicate the authors' hypotheses rather than results from literature.

ances that were parameterised to project two extreme traits. On average, the utterances generated with PERSONAGE-PE were found to be significantly less natural than those generated with PERSONAGE.

Compared to Mairesse and Walker's work, our models for the generation of language that exhibits individual differences, ALPM-3 and ALPM-4 (Chapters 4 and 6), follow an approach of mild overgeneration of paraphrases with subsequent re-ranking according to n-gram language models, as supported by the OPENCCG surface realiser (White 2006b). Mairesse and Walker (2008) argue that this paradigm is detrimental to real-time generation. OPENCCG employs n-gram models in a best-first anytime search, in such a way that preferred realisations tend to be found early in the search process, so its performance can be influenced by setting time limits. However, we did not investigate this issue further because the materials for the experiments reported in this thesis were generated offline.

In contrast to PERSONAGE, which modifies the personality of individual utterances in the restaurant recommendation domain, ALPM-4 generates movie review dialogues between computer characters that vary in personality and, additionally, alignment behaviour. This unified treatment of personality and alignment in dialogue is a novel contribution.

2.4.3 Stylistically Controlled Generation

Other related work aimed at modifying the *style* of generated language. Hovy (1988) developed the Planning And Uttering Language In Natural Environments (PAULINE) program, which generated text that varied according to pragmatic constraints. The system was based on a set of pragmatic features like the interlocutors' knowledge or opinions of the topic, or their emotional state. Intermediate rhetorical strategies were used to link the pragmatic features to generator decisions.

Paiva and Evans (2004) determined stylistic dimensions from a corpus of texts via factor analysis. After that, a number of texts in the domain of the corpus were generated. The generator's decisions at key choice points and style scores for the resulting text were recorded and subsequently correlated. The correlations were then used to influence the style of subsequent generation.

The models developed in the course of this thesis do not take higher-level pragmatic or stylistic features into consideration; they are restricted to lexically-based manipulations. The idea is to see how far these approaches can take us towards recognisable

differences in personality and alignment behaviour.

2.4.4 Human–Computer Interaction

Previous research on computer interfaces has shown that people treat computers as if they were people. In particular, there is evidence that computer users attribute personality to interfaces, and respond to it in robust ways (Reeves and Nass 1996, Isbister and Nass 2000, Nass and Lee 2001). Even in a text-only environment, extraverts preferred interfaces presenting information using language associated with extravert traits; introverts preferred introverted interfaces. An interface with matching personality was judged more positively, and rated as more attractive, credible and informative.

In addition to these results on personality, alignment-related effects were confirmed in human–computer interaction as well. Brennan (1996) studied *lexical entrainment*, the phenomenon that interlocutors agree on referring expressions in the course of a conversation. People converged not only in dialogues with human partners, but also adopted terms used by natural language computer interfaces. Pearson et al. (2006) observed that users adapted their language according to their expectations about a computer they were interacting with. In a picture-naming and -matching experiment, the computer was modified to display a basic or an advanced start-up screen, while it behaved identically otherwise. Participants aligned, i.e., used the same term as the computer to describe an object, significantly more than chance in both conditions. However, alignment was greater when participants thought they interacted with the basic computer.

2.4.5 Generation of Dialogues between Computer Characters

In this thesis, we are not concerned with dialogues between humans and computers, but instead we generate dialogues between two computer characters which are then presented to human judges. Related work in the area of embodied conversational agents (ECAs) is concerned with the implementation of computer characters that exhibit human behaviour. André et al. (2000) describe experiments with teams of animated presentation agents with scripted behaviour. One of the systems generates animated dialogues in the car sales domain. The agents' level of extraversion and agreeableness and their interests can be configured in advance and have effects

on their dialogue strategies; for example, extravert agents start the conversation. The generation is template-based.

In the Net Environment for Embodied Emotional Conversational Agents (NECA) framework (Piwek 2003, van Deemter et al. 2008), dialogues between ECAs are *fully* generated according to abstract scripts. The ESHOWROOM demonstrator, a NECA application, simulates a car sales dialogue between a seller and a buyer, similar to André et al. (2000)'s system described above. A user can specify preferences about cars, e.g., *price* or *environmental friendliness*, and influence the ECA's personality by choosing a combination of *polite* or *impolite* and *good humored* or *ill tempered* settings. The system then incrementally builds the specification of a multimodal dialogue in NECA's Rich Representation Language (RRL), which combines information about utterances' content, their textual string of words, and about speech and body language. The final specification is then converted and presented to the user by an animation player. In an evaluation, users found the demonstrator enjoyable, and judged that the initially configured preferences matched the system's output.

The implementations of the Affective Language Production Models developed within this thesis produce movie review dialogues between two computer characters. As in the related work, the characters' personality and alignment behaviour as well as their topic agenda and opinions are configured in advance. However, we limit ourselves to the study of textual output and do not attempt to generate animated movies with ECAs. ALPM-1 and ALPM-2 (Chapter 3) create dialogues by re-combining existing utterances according to their personality rating, while ALPM-3 and ALPM-4 (Chapters 4 and 6) fully generate utterances from logical forms.

Chapter 3

Recognising Personality from Text

3.1 Summary

Previous research on the relationship between personality traits and language identified a set of features associating the three personality dimensions extraversion (E), neuroticism (N) and psychoticism (P) with specific language use (Gill and Oberlander 2002, Gill 2004, Oberlander and Gill 2004, 2006). This chapter develops a model that uses these features to rank arbitrary unseen utterances by personality. The model operationalises the view that a simple computer program can simulate (aspects of) an individual's personality-based language choices. For evaluation purposes, a corpus of movie review dialogues was compiled from recorded conversations of people discussing a movie they had seen. Their utterances were then ranked by personality and re-combined to produce dialogues between two computer characters, according to Oberlander and Gill's Affective Language Production Model (ALPM-1 and ALPM-2), while systematically varying the personality parameters. The dialogues were presented to human judges in order to assess whether they could reliably perceive personality differences. It was found that extraversion could be detected according to expectations, while neuroticism was more difficult to recognise. Psychoticism was perceived contrary to our initial hypothesis that tough-minded characters stick to their own topic agenda. The results held for both native and non-native speakers of English.

The research discussed in this chapter was conducted in collaboration with colleagues from the Critical Agent Dialogue project (CrAg). The CrAg Corpus was collected, transcribed and manually annotated by Amy Isard; the annotation was automatically augmented by components implemented by Carsten Brockmann. The utterance ranking approach was developed by Carsten Brockmann, Amy Isard and Jon Ober-

lander and implemented and evaluated by Carsten Brockmann. The corpus and the approach were described in Isard et al. (2005).

3.2 Introduction

The motivation for the model described in this chapter stems from previous research on the perception of personality from text. Correlations were found between participants' personality dimensions and their language use, and corresponding features were identified. The hypothesis examined here is that these features can be employed to choose between alternative textual realisations, as a first step on the way leading to natural language generation influenced by personality.

The approach taken to verify the hypothesis is to create a model that can be parameterised for personality. Then dialogues between computer characters are assembled by choosing the utterances that best fit the respective parameter settings. These dialogues are presented to human judges in order to find out whether they can perceive differences in personality.

Gill et al. (2006) studied the perception of personality in asynchronous computer-mediated communication using Eysenck's three-factor model. They characterised the dimensions in terms of *visibility* and *evaluativeness*. According to this approach, E would be most easily perceived because it is highly visible and low in evaluativeness. Both N and P are less visible and more evaluative, and thus more difficult to perceive.

In their experiment, Gill et al. found that extraversion and to a lesser extent psychoticism could be perceived by judges in a task involving exemplar-based rating of personality. Judges agreed both with each other and with the target individual's self-rating. For neuroticism, judges agreed with each other, but their judgements did not match the self-assessments.

3.2.1 Hypotheses

Our hypotheses relating language behaviour to implementable computational features are listed in full in Section 3.3.2, after the required methods and resources have been introduced. We expect the results of the perception experiment to agree with previous studies on the perception of personality at zero acquaintance.

3.3 Methods

3.3.1 CrAg Corpus

The Critical Agent Dialogue project (CrAg) Corpus was built in order to obtain a data source in the domain of movie reviews, with speakers of known personality.

3.3.1.1 Collection

Ten pairs of participants went to see a film that was chosen for them and were later recorded having a conversation about it. Three films were selected which were showing at the same time, and were from three different genres: Intolerable Cruelty (IC; romantic comedy), League of Extraordinary Gentlemen (LXG; action, science fiction, fantasy) and Mystic River (MR; drama, crime). The dialogues were recorded in a soundproof room. Participants were told that they could talk about any aspect of the film they had just watched, and asked to try to stay on that film as their topic, but the conversation was not monitored. The dialogues ranged in length from 12 to 25 minutes, with an average of 19 minutes.

The participants also completed questionnaires to determine their personality according to the five factor model (Costa and McCrae 1992), so that their language use could later be associated with their personality.

3.3.1.2 Transcription and Annotation

3.3.1.2.1 Topic and Sentiment Polarity The dialogues were segmented into phrases and transcribed orthographically. The corpus was manually scanned for recurring discussion topics, which yielded a list of 12 topics plus four further topic-like categories, as shown in Table 3.1. A single annotator assigned one or more topics from this list to each utterance, and also one of the sentiment polarities displayed in Table 3.2.

This process resulted in a total of 1465 utterances, averaging 73 per speaker. The topics were not distributed evenly throughout the dialogues since the films belonged to three different genres, and some topics (e.g., SPECIAL EFFECTS) did not apply to all types of film. The exact frequencies are listed in Table 3.3.

3.3.1.2.2 Generality With re-generation in mind, utterances were labelled according to whether they made sense out of context. This meant that most utterances

Topic	Definition
ACTION SEQUENCES	
ACTORS	
CHARACTERS	
CINEMATOGRAPHY STYLE	The look of the film.
DIALOGUE	
DIRECTING	Directing style, director's intentions etc.
HUMOUR	
MUSIC	
ROMANCE	
SPECIAL EFFECTS	
STORY	
WHOLE MOVIE	
OTHER (THIS FILM)	A topic related to this film not included in the above list.
OTHER FILM	A discussion about another film or films.
NOT FILM-RELATED	Any discussion not related to films at all.
NONE	Utterances where a topic could not be assigned, e.g., *um*, *he it*, *I . . . I think*.

Table 3.1: Topics assigned to the utterances in the CrAg Corpus. Topics without a definition are assumed to be self-explanatory.

Polarity	Definition
POSITIVE	The speaker liked the topic under discussion.
NEGATIVE	The speaker disliked the topic under discussion.
MIXED	Both positive and negative, e.g., *All the cinematography was alright there was nothing interesting in it nothing daring.*
UNCLEAR	Neutral, e.g., *What did you think of Sean Connery?*, *Well there's a clear implication that they had a relationship before.*
	Not possible to assign a polarity; UNCLEAR was automatically assigned to all utterances whose topic was NONE.

Table 3.2: Sentiment polarities assigned to the utterances in the CrAg Corpus.

	Film			
Topic	IC	LXG	MR	All
ACTION SEQUENCES	0	11	16	27
ACTORS	66	30	95	171
CHARACTERS	52	110	282	444
CINEMATOGRAPHY STYLE	0	7	12	19
DIALOGUE	8	37	8	53
DIRECTING	48	23	65	136
HUMOUR	76	5	2	83
MUSIC	0	0	25	25
ROMANCE	9	0	8	17
SPECIAL EFFECTS	0	48	0	48
STORY	83	165	245	493
WHOLE MOVIE	36	74	44	154
OTHER	173	106	124	403
Total	401	427	637	1465

Table 3.3: CrAg Corpus overall utterance counts by topic. (N.B., because there can be more than one topic per utterance, the totals at the bottom are less than the sum of their columns.)

	Film				
Topic	IC	LXG	MR	General	All
ACTION SEQUENCES	0	6	1	0	7
ACTORS	4	1	12	2	19
CHARACTERS	1	7	12	3	23
CINEMATOGRAPHY STYLE	0	5	2	2	9
DIALOGUE	2	3	0	4	9
DIRECTING	3	1	2	1	7
HUMOUR	8	2	0	0	10
MUSIC	0	0	1	0	1
ROMANCE	0	0	0	0	0
SPECIAL EFFECTS	0	14	0	0	14
STORY	2	14	12	7	35
WHOLE MOVIE	4	10	11	19	44
Total	15	44	41	32	132

Table 3.4: CrAg Corpus context-independent utterance counts by topic.

with anaphoric references were rejected (e.g., *There was no mention of that at all in the film.*) along with those which required knowledge of the previous utterance (e.g., *And there wasn't even that much blood-sucking which is kind of disappointing for a vampire.*) and questions (e.g., *What did you think?*).

For the same reason, the utterances were also annotated for whether they could apply to just one film (e.g., *They'd start little storylines like when Sean Connery was teaching the American chap to shoot.*) or could be used to discuss any film (e.g., *I don't have anything positive to say about it actually.*).

Table 3.4 shows counts of the context-independent utterances which were considered to be usable for re-generation. Utterances listed under each film are those which could only be used in a discussion of that particular film, and those in the *General* column were applicable for any film.

3.3.1.3 Augmenting the Annotation

The corpus utterances' manual annotation was augmented with information from a variety of machine-readable linguistic and psycholinguistic resources. This knowledge was then used to compute E and N scores (see Section 3.3.2).

3.3.1.3.1 Part-of-Speech Tagging and Lemmatisation Each utterance was split into sentences, tokenised and tagged with part-of-speech (POS) information using the Maximum entropy model for Part-Of-Speech Tagging (MXPOST, Ratnaparkhi 1996). The MORPH tool (Minnen et al. 2001) then determined each word's lemma form.

Based on the lemmata, each utterance's type/token ratio was computed, which measures the variety of words used; it equals 1 if every type is used only once, and decreases with each repetition.

3.3.1.3.2 MRC Psycholinguistic Database The annotation was further augmented by information from the MRC Psycholinguistic Database (MRCPD, Wilson 1988), a machine readable dictionary of 150,837 words. The MRCPD was derived from a database that was originally produced under a grant from the Medical Research Council (MRC). For each word, it specifies up to 26 linguistic and psycholinguistic attributes, e.g.:

- Written/spoken word frequencies.
- Familiarity, concreteness, imageability.
- Meaningfulness.
- Age of acquisition.
- Part of speech.
- Phonetic transcription, stress pattern.

3.3.1.3.3 Linguistic Inquiry and Word Count Linguistic Inquiry and Word Count (LIWC, Pennebaker et al. 2001) is another machine readable dictionary. 2,300 words and word stems are annotated with one or more of 74 categories, e.g.:

- Linguistic dimensions (pronouns, negations, articles, . . .).

- Psychological processes:
  - Positive/negative emotions.
  - Cognitive processes (insight, certainty, ...).
  - Perceptual processes (seeing, hearing, feeling).
  - Social processes (friends, family, ...).
- Relativity (time, space, motion).
- Personal concerns (occupation, leisure, physical states, ...).

3.3.1.3.4 The Formality Measure F The formality measure F (Heylighen and Dewaele 2002) is computed from each utterance's part of speech annotation. The authors propose the concept of formality as a "dimension of variation between linguistic expressions". The measure is based on frequency percentages of different word classes:

$$\begin{aligned} F = &(\text{noun freq.} + \text{adjective freq.} + \text{preposition freq.} \\ &+ \text{article freq.} - \text{pronoun freq.} - \text{verb freq.} \\ &- \text{adverb freq.} - \text{interjection freq.} + 100)/2 \end{aligned} \tag{3.1}$$

In Heylighen and Dewaele's study, oral female ($F = 38.7$) and oral male ($F = 41.6$) language was classified as informal; novels ($F = 52.5$) were average, while scientific text ($F = 65.7$) and newspapers ($F = 68.1$) ranked high on the formality scale.

3.3.2 Feature Combination

Previous research identified features characteristic of the language of extravert or neurotic speakers (Pennebaker and King 1999, Gill and Oberlander 2002, Gill 2004, Oberlander and Gill 2004, 2006). According to these results, the utterance scores computed during the annotation phase are combined using additive multiattribute value functions (AMVFs). AMVFs have been applied to represent user preferences (Carenini and Moore 2000); an implementation developed for the user modelling component of the Fancy Linguistically Informed Generation of Highly Tailored Speech system (FLIGHTS, Moore et al. 2004) is being re-used here.

In an AMVF, a value tree specifies the hierarchy of aspects of an entity e. Edges are weighted (w) according to the importance of their contribution to the parent node.

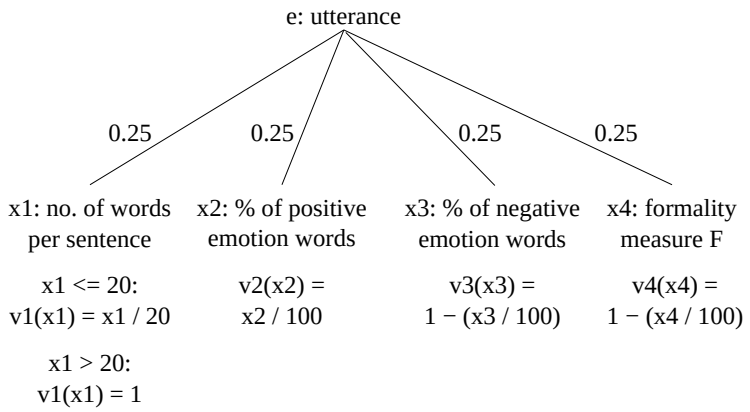

Figure 3.1: Partial additive multiattribute value function (AMVF) for extravert language.

For each leaf, a component value function v_i maps attribute value x_i to the $[0,1]$ interval (1 is most preferable). The weight w_i of a leaf node is computed as the product of the weights from the tree's root down to the leaf. Given this model, the value $v(e)$ of entity e can be computed:

$$v(e) = v(x_1, \ldots, x_n) = \sum_{i=1}^{n} w_i v_i(x_i) \tag{3.2}$$

In FLIGHTS, attributes may be ranked in order of importance. A partial ordering of rankings is supported, i.e., several attributes can be given equal importance. Ranks are converted into attribute weights (Moore et al. 2004, p. 3):

> [...] As in previous work, we use Rank Order Centroid (ROC) weights (Edwards and Barron 1994). This allows weights to be assigned based on rankings, guaranteeing that the sum will be 1. The n^{th} ROC weight w_n^R of N total weights is computed as follows: $w_n^R = \frac{1}{N}\sum_{i=n}^{N}\frac{1}{i}$.
>
> We extend these initial weights to the partial-ordering case as follows. If attributes $i \ldots j$ all have the same ranking, then the weight of each will be the mean of the relevant ROC weights; that is, $(\sum_{k=i}^{j} w_k^R)/(j-i+1)$. As a concrete example, if there is a single highest-ranked attribute followed by a three-way tie for second, then $w_1 = w_1^R$, while $w_2 = w_3 = w_4 = \frac{1}{3}(w_2^R + w_3^R + w_4^R)$.

A simplified example AMVF for extravert utterances is shown in Figure 3.1. The complete set of ranked features characteristic of high extravert language is listed in Table 3.5, and the ranked features for high neurotic language can be found in Table 3.6.

The tables also specify the component value functions corresponding to each feature. In general, if a feature is expected to be high, its actual value is divided by its

maximum value in order to map it to the $[0, 1]$ interval. If a feature is expected to be low, this number is subtracted from 1. For instance, the maximum value of a percentage is 100; the maximum value of the Brown verbal frequency taken from the MRCPD is 6833. For the number of words per sentence and the number of sentences per utterance, 20 and 5 were set as the maximum values, respectively. Values exceeding these limits are mapped to 1.

Mairesse et al. used machine learning approaches to find optimal feature sets and weights for personality recognition (Mairesse and Walker 2006a,b, Mairesse et al. 2007). In the present model, the features were selected and ranked manually according to previous research results in order to investigate whether this simpler approach is sufficient. Simplicity of personality models has also been advocated in related research on human–computer interfaces by Reeves and Nass (1996, pp. 90 ff.), who studied computer personality in terms of dominance and submissiveness. While we keep our model simple in this chapter, ALPM-3 and ALPM-4 in later chapters will be more sophisticated.

3.3.3 Re-Generating Dialogue

3.3.3.1 Initialisation

Computer characters are defined by values for the personality dimensions E, N, and P. These values are given in a range from 0 (low) to 1 (high). For P, only the two settings low ($P < 0.5$) and high ($P \geq 0.5$) are distinguished, as explained below. The characters are also each assigned an agenda of topics about which they would like to talk; for each topic, their opinion about it (the *polarity*) is either positive or negative.

Dialogues between two computer characters are then re-generated by the Open Agent Architecture (OAA, Martin et al. 1999) Critical Agent Dialogue system version 1 (CRAG 1) Steering Agent. Two character definitions and one of the three available films are selected, and the number of turns to generate is set.

3.3.3.2 The Affective Language Production Model

The generation process is informed by the Affective Language Production Model (ALPM), developed by Oberlander and Gill. The simplest version of this model (ALPM-1) starts from the idea that high extraverts have plenty of resource for linguistic interaction, but need to put less of it into detailed planning. High neurotics have less resource for linguistic interaction in the first place. It follows that extraversion

Rank	Level	Feature, Component Value Function	Examples	Resource
1	High	Number of words per sentence		
		$x \leq 20$: $v(x) = x/20$; $x > 20$: $v(x) = 1$		
		Number of sentences per utterance		
		$x \leq 5$: $v(x) = x/5$; $x > 5$: $v(x) = 1$		
		Percentage of positive emotion words	*happy*, *pretty*, *good*	LIWC
		$v(x) = x/100$		
		Percentage of social process words	*talk*, *us*, *friend*	LIWC
	Low	Percentage of tentative words	*maybe*, *perhaps*, *guess*	LIWC
		$v(x) = 1 - (x/100)$		
2	High	Mean Brown verbal frequency		MRCPD
		$v(x) = x/6833$		
		Percentage of inclusive words	*with*, *and*, *include*	LIWC
	Low	Percentage of negations	*no*, *never*, *not*	LIWC
		Percentage of negative emotion words	*hate*, *worthless*, *enemy*	LIWC
		Percentage of causation words	*because*, *effect*, *hence*	LIWC
		Percentage of discrepancy words	*should*, *would*, *could*	LIWC
		Formality		F measure
		$v(x) = 1 - (x/100)$		
3	High	Percentage of conjunctions		MXPOST
		Percentage of pronouns		MXPOST
		Percentage of first person sing. pron.	*I*, *my*, *me*	LIWC
		Percentage of certainty words	*always*, *never*	LIWC
		Percentage of present tense verbs	*walk*, *is*, *be*	LIWC
	Low	Percentage of prepositions		MXPOST
		Percentage of exclusive words	*but*, *except*, *without*	LIWC

Table 3.5: Features characteristic of high extravert language, with corresponding component value functions. The two component value functions for high or low percentages of features, respectively, do not differ across features and are therefore only specified at their first occurrence.

Rank	Level	Feature, Component Value Function	Examples	Resource
1	High	Percentage of first person sing. pron.	*I*, *my*, *me*	LIWC
		$v(x) = x/100$		
		Percentage of negative emotion words	*hate*, *worthless*, *enemy*	LIWC
	Low	Percentage of positive emotion words	*happy*, *pretty*, *good*	LIWC
		$v(x) = 1 - (x/100)$		
2	High	Mean Brown verbal frequency		MRCPD
		$v(x) = x/6833$		
		Percentage of discrepancy words	*should*, *would*, *could*	LIWC
	Low	Percentage of nouns		MXPOST
		Percentage of articles	*a*, *an*, *the*	LIWC
		Formality		F measure
		$v(x) = 1 - (x/100)$		
3	High	Percentage of adjectives		MXPOST
		Percentage of conjunctions		MXPOST
		Percentage of verbs		MXPOST
		Percentage of negations	*no*, *never*, *not*	LIWC
		Percentage of tentative words	*maybe*, *perhaps*, *guess*	LIWC
		Percentage of present tense verbs	*walk*, *is*, *be*	LIWC
	Low	Percentage of prepositions		MXPOST
		Lemma-based type/token ratio		MORPH
		$v(x) = 1 - x$		
		Percentage of social process words	*talk*, *us*, *friend*	LIWC

Table 3.6: Features characteristic of high neurotic language, with corresponding component value functions. The two component value functions for high or low percentages of features, respectively, do not differ across features and are therefore only specified at their first occurrence.

finds its effects mostly at the stages of formulation (surface realisation). That is, the process and representations used in realisation differ between high and low extraverts. Secondly, neuroticism finds its effects at the stage of conceptualisation (content selection). That is, the process and representations used in content selection differ between high and low neurotics. Since content selection precedes surface realisation, variations in neuroticism will have consequences beyond the content selection stage, but this is their primary locus.

ALPM-2 adds to ALPM-1 an intuitive treatment of psychoticism. The idea is that some individuals pay more attention to their partner's utterances than others. Those who pay attention are more likely to produce utterances which overlap in both content (and form) with those of their partner.

3.3.3.3 Re-Generation

Thus, the re-generation process operationalises ALPM-2 in the following way.

First, the character with the higher extraversion score begins the dialogue, and the topic at the top of their agenda is used.

Subsequent topic choice is influenced by the characters' psychoticism score. A low psychotic character continues on the same topic as the previous speaker, while a high psychotic changes topic to the one at the top of their own agenda.

Next the corpus is queried for all utterances with the appropriate film, topic, and polarity. By the previous annotation process (cf. Section 3.3.1.3) each utterance was assigned an N and an E score.

A character's N level relates to content selection. Thus, the N score is the first filter that is applied to rank the utterance candidates. To implement this as simply as possible, the half of the candidates which most closely match the desired N score is kept, while the other half is discarded.

A character's E level relates to surface realisation behaviour. So the remaining candidates are sorted by E score, and the one that is closest to the character's desired E score and has not been used in the preceding dialogue is chosen as their actual utterance.

After this, it is the other character's turn to say something. The routine loops back to the point of topic choice.

	Film					
	IC		LXG		MR	
Topic	+	−	+	−	+	−
ACTORS	6	4	0	3	11	5
CHARACTERS	0	3	0	9	6	10
CINEMATOGRAPHY STYLE	0	0	5	0	0	0
DIRECTING	0	5	0	5	0	6
HUMOUR	5	3	0	0	0	0
SPECIAL EFFECTS	0	0	0	7	0	0
STORY	0	9	0	25	0	19
WHOLE MOVIE	3	23	4	29	4	26

Table 3.7: CrAg Corpus utterances available for re-generation.

3.3.4 Materials

3.3.4.1 Generation

In order to prepare an evaluation of the approach, the topic annotation of the CrAg Corpus was revised. The number of topics was reduced from 12 to 8: The topics ACTION SEQUENCES, DIALOGUE, MUSIC and ROMANCE were removed due to low overall and context-independent utterance counts in the original annotation. Two annotators went over all utterances again and assigned topics, then conflicts were resolved by the first annotator. General (film-independent) utterances were allowed to be used in each individual movie so that enough choice was available for the re-generation of varying dialogues. Table 3.7 shows the updated figures.

For all of the resulting 147 utterances, scores were computed for the E and the N dimension. Figure 3.2 shows the distribution of the scores. E scores ranged from 0.3953 to 0.5399, with mean $\bar{x} = 0.4901$ and standard deviation $s = 0.0341$. N scores ranged from 0.2572 to 0.4780, with $\bar{x} = 0.3907$ and $s = 0.0295$. The mean scores were used as the neutral or middle setting when generating the experimental materials.

Dialogues were generated in four different conditions, as shown in Table 3.8. Each condition sets the two computer characters to opposing extremes on either the E or the N dimension, while keeping the respective other dimension at a middle level. Further-

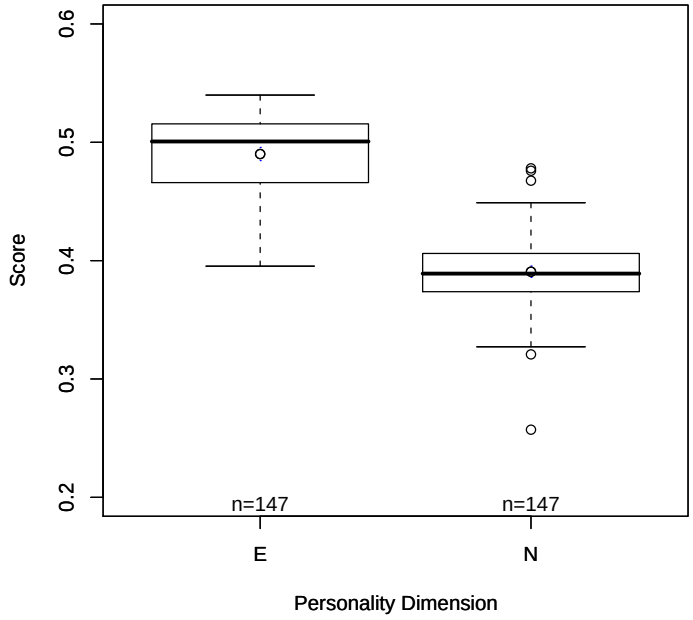

Figure 3.2: Boxplot for the personality scores computed for the CrAg Corpus utterances available for re-generation.

more, character A is always High P, and character B is always Low P.

Four dialogues were generated per condition. The movie under discussion and the characters' agenda and their opinions about the topics were randomly assigned. Each dialogue was eight utterances long, with characters taking turns, each of them producing four utterances altogether.

3.3.4.2 Web Experiment

The experiment was conducted on the World Wide Web, advertised via the Language Experiments Portal[1]. An introductory page explained the task and defined the three personality dimensions by paraphrasing descriptions from Eysenck and Eysenck (1975). The technical terms *neuroticism* and *psychoticism* were replaced by *emotional instability* and *tough-mindedness*, respectively, to make them more accessible to laymen and to avoid or reduce evaluativeness.

Extraversion Typical extraverts are sociable, like parties, have many friends, need to

[1] `http://www.language-experiments.org/`

		Personality Parameter Setting		
Condition	Character	E	N	P
A) High E	A	1	0.3907	1
vs. Low E	B	0	0.3907	0
B) Low E	A	0	0.3907	1
vs. High E	B	1	0.3907	0
C) High N	A	0.4901	1	1
vs. Low N	B	0.4901	0	0
D) Low N	A	0.4901	0	1
vs. High N	B	0.4901	1	0

Table 3.8: Personality parameter settings for the four experimental conditions.

have people to talk to, and do not like reading or studying by themselves. They crave excitement, take chances, often stick their neck out, act on the spur of the moment, and are generally impulsive individuals. They are fond of practical jokes, always have a ready answer, and generally like change; they are carefree, easy-going, optimistic, and like to "laugh and be merry." They prefer to keep moving and doing things, tend to be aggressive and lose their temper quickly; altogether their feelings are not kept under tight control, and they are not always reliable people.

Emotional instability Typical emotionally unstable people are anxious, worrying individuals, moody and frequently depressed. They are likely to sleep badly, and to suffer from various psychosomatic disorders. They are overly emotional, reacting too strongly to all sorts of stimuli, and find it difficult to get back on an even keel after each emotionally arousing experience. Their strong emotional reactions interfere with their proper adjustment, making them react in irrational, sometimes rigid ways. If emotionally unstable individuals have to be described in one word, one might say that they are *worriers*; their main characteristic is a constant preoccupation with things that might go wrong, and a strong emotional

Dimension	Exemplars
Extraverted	Sociable, talkative, impulsive, carefree, optimistic.
Emotionally unstable	Anxious, worrying, moody, strongly emotional, irrational.
Tough-minded	Solitary, uncaring, insensitive, aggressive, manipulative.

Table 3.9: Exemplars for the personality dimensions displayed during the web experiment.

reaction of anxiety to these thoughts.

Tough-mindedness Tough-minded individuals may be described as being solitary, not caring for people; they are often troublesome, not fitting in anywhere. They may be cruel and inhumane, lacking in feeling and empathy, and altogether insensitive. They are hostile to others, even with their own kith and kin, and aggressive even to loved ones. They have a liking for odd and unusual things, and a disregard for danger; they like to make fools of other people and to upset them.

The introductory page also contained an example dialogue similar to those that would be presented during the experiment. The complete instructions are reproduced in Section A.1 in the appendix.

Subsequently, the participants were presented with two dialogues per condition, randomly chosen from the pool of available dialogues, with the constraint that a dialogue from the A or B condition had to be followed by one from the C or D condition, and vice versa. One dialogue was shown at a time, and the participants were asked to judge which of the two characters scored higher on the E, N and P dimensions. The higher-scoring character had to be chosen; *equal* was not an option. Five exemplars for each personality dimension were displayed at the bottom of the page (see Table 3.9). The participants were asked to rate each dialogue independently from the others.

The experiment was open to native as well as non-native speakers of English to allow an analysis of whether the results held for both groups. Participants were entered into a draw for a £15 Amazon.co.uk gift certificate if they chose to specify a valid e-mail address. All data were analysed anonymously.

3.3.4.3 Example Dialogues

To give an impression of what the generated dialogues look like, this section presents one dialogue each of the A (High E versus Low E) and C (High N versus Low N) conditions. The parameter settings and movie and topic information were not shown to the experiment's participants.

Some of the features can be described relatively easily. In condition A, the High E character A's utterances are usually longer than those of character B. In condition C, the High N character A uses the first person singular word *I* more often (5 times) than character B (once). In both conditions, the High P character's strategy of switching topics in order to follow his own agenda is obviously recognisable.

3.3.4.3.1 High E versus Low E

Parameter Settings

Character A's personality High E, Mid N, High P

Character B's personality Low E, Mid N, Low P

Movie League of Extraordinary Gentlemen

Dialogue

A (CINEMATOGRAPHY STYLE) "It certainly wasn't as visually exciting as X-Men I didn't think, it was quite fun, it worked."

B (CINEMATOGRAPHY STYLE) "It was colourful."

A (STORY) "Even they'd start little storylines, like when Sean Connery was teaching the American chap to shoot and he made comments because they knew that his son had been killed, and the chap made a comment to Sean Connery about his son, and then Sean Connery leaves and so you're left feeling this is a troubled man, but it's never developed, and there's never a conversation about it again it's never even referred to."

B (STORY) "It wasn't nicely tied up."

A (SPECIAL EFFECTS) "They could have just gone normally, I felt the ship should have been a wooden ship, and they went all for the glossy shiny robotic things going on."

B (SPECIAL EFFECTS) "It's almost as if they had used or looked at various other films and decided they would take snippets of the technology from each, because for instance Mr Hyde, when he was first seen fighting or whatever, that reminded me so much of that monster in Lord of the Rings."

A (CHARACTERS) "Again if you go back to the comic it's better thought out, and again the the character interaction is better thought out, there are various points of real tension and especially things like keeping Mr Hyde in check, he's a monster and he has evil appetites, he rips people apart, he's a violent evil creature."

B (CHARACTERS) "I don't know it seemed to have aspects of lots of different films, a bit of suspense characters with a few strange abilities, but it didn't quite hit the spot with any of them really, it was a bit of a mishmash in between."

3.3.4.3.2 High N versus Low N

Parameter Settings

Character A's personality Mid E, High N, High P

Character B's personality Mid E, Low N, Low P

Movie Mystic River

Dialogue

A (ACTORS) "It had no relation to real life which is fair enough for some films, but not if you're actually going for something of a real life film."

B (ACTORS) "They were acting up to the part, they didn't act like real people, they were acting like somebody in a Hollywood movie is."

A (STORY) "It could have been a better cop movie or something I guess, I mean focussing on Sean Penn and and him going after revenge wasn't that interesting I guess."

B (STORY) "It might have made a really good book or something, maybe with a good book you have enough time to elaborate on on the actual symbolism behind all of it."

A (DIRECTING) "The director chooses what they're going to film, and what kind of slant to take on it, maybe he just took too much of everything rather than going for one, maybe if it had one driving vision and one character it would have worked."

B (DIRECTING) "All the cinematography was alright, there was nothing interesting in it, there was nothing daring, in the same way the direction and editing, there was nothing new, there was nothing to keep you hooked on it, it was just kind of presenting a story."

A (WHOLE MOVIE) "I thought it could have ended a lot sooner, just before the parade bit I think, just before that."

B (WHOLE MOVIE) "I can think of very few positive things to say about it really."

3.4 Results

Ninety-two participants completed the web experiment. The data were filtered to exclude replies that were submitted after less than five minutes (18 cases) or more than 45 minutes (three cases), leaving 71 judgements for further analysis. Of these, 41 were native and 30 were non-native speakers of English.

The data were evaluated by comparing the settings used to generate the experimental materials with the actual human judgements. The hypothesis was that for conditions A and B people would be able to detect differences in extraversion, while for conditions C and D they would see differences in neuroticism. Differences in psychoticism were expected across all four conditions. Table 3.10 shows the results of this comparison. Figures 3.3 to 3.5 show the percentage of participants' choices in agreement with expectations for native, non-native and all participants, respectively.

Binomial tests were performed to assess whether the distribution of values differed significantly from chance. Significance levels were stable across all three categories. Participants significantly agreed with the model in their judgement of extraversion (native, non-native and all: $p \leq 0.001$). The neuroticism judgements tended towards expected choices but never reached significance. The psychoticism dimension was significantly judged not as expected (native: $p \leq 0.01$; non-native: $p \leq 0.05$; all: $p \leq 0.001$).

	Native Speakers			Non-Native Speakers			All		
Choice	E	N	P	E	N	P	E	N	P
Expected	110	85	140	90	65	100	200	150	240
Not expected	54	79	188	30	55	140	84	134	328
Number of trials	164	164	328	120	120	240	284	284	568
Expected (%)	67.1	51.8	42.7	75.0	54.2	41.7	70.4	52.8	42.3
Significance	$***$		$**$	$***$		$*$	$***$		$***$

Table 3.10: Numbers of expected versus not expected choices for the web experiment. Significance levels of binomial test: $*$: $p \leq 0.05$, $**$: $p \leq 0.01$, $***$: $p \leq 0.001$.

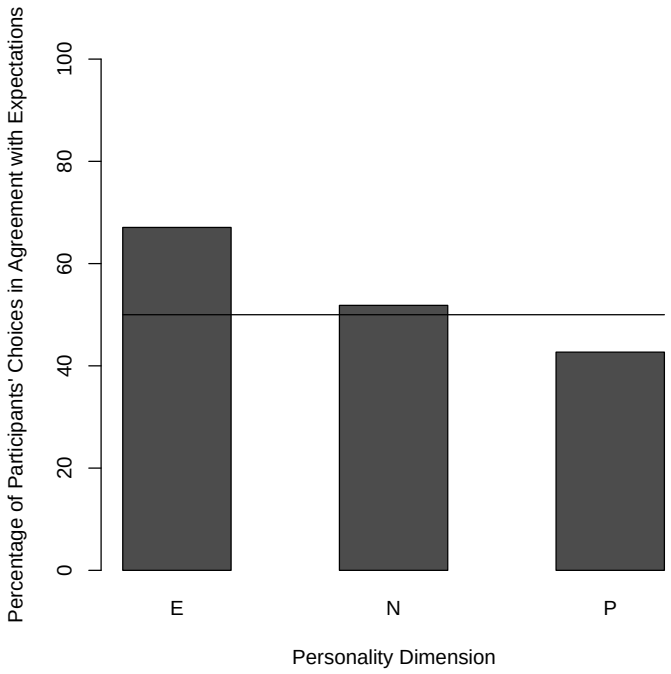

Figure 3.3: Percentage of native English speakers' choices in agreement with expectations, by personality dimension.

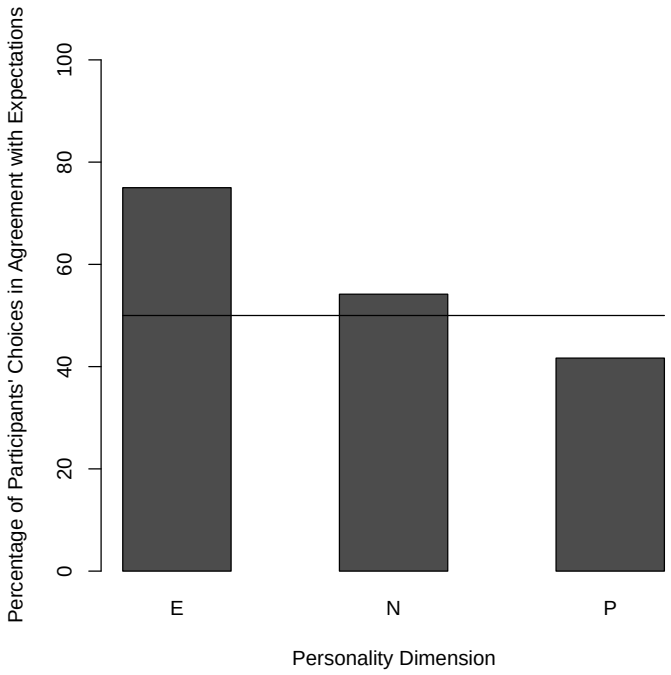

Figure 3.4: Percentage of non-native English speakers' choices in agreement with expectations, by personality dimension.

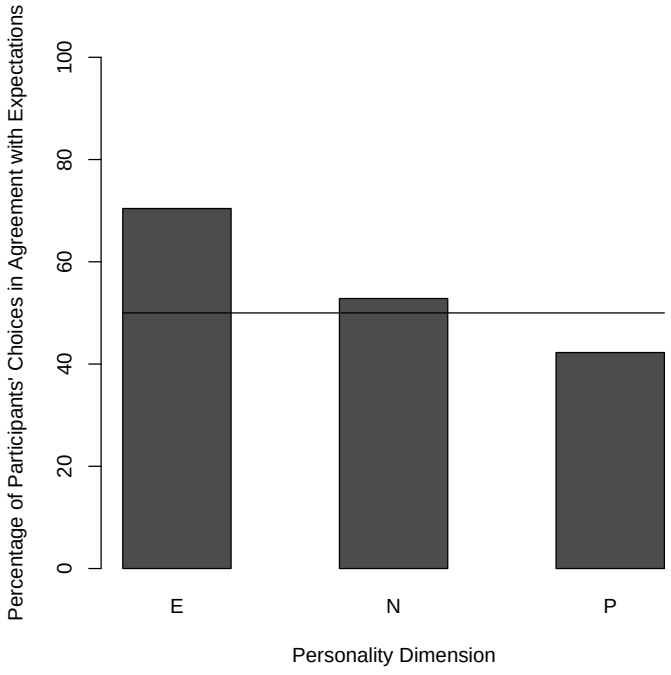

Figure 3.5: Percentage of all participants' choices in agreement with expectations, by personality dimension.

3.5 Discussion

Participants agreed with ALPM-2 in their judgement of extraversion. This suggests that the features used to model this dimension are adequate. It is also in agreement with Gill et al.'s (2006) finding that judges agreed on E ratings ($r_s = 0.482$) and were able to significantly identify E levels in accordance with self-ratings ($r_s = 0.886^*$).

The trend for the neuroticism dimension leans toward ALPM-2's expected direction, but the results do not differ significantly from chance. This can be explained by N's lower visibility and higher evaluativeness. Consistent with this, Gill et al. found inter-judge agreement for N ($r_s = 0.308$) lower than for E; the target-judge agreement ($r_s = -0.377$) was even non-significantly negative.

Results for the psychoticism dimension contradict our initial intuitive ALPM-2 hypothesis that tough-minded characters stick to their own agenda and switch topics instead of sticking to the one that their dialogue partner used. Judges perceived this exactly the other way around: With high significance, they assumed that characters who stick to the other's topic are tough-minded. This indicates that our hypothesis should be rejected. Gill et al. found inter-judge agreement for P ($r_s = 0.333$) slightly higher than that for N and relatively high target-judge agreement ($r_s = 0.754$) – in our case, the judges' choices were highly significantly different from chance, comparable to those for E, so if we revise the direction of our initial hypothesis, it fits those results.

This suggests that tough-mindedness may be perceived as willingness to engage with another's opinion (to either agree or disagree with it). Failure to do so – as when an agent changes topic – appears to be perceived as being the opposite of tough-minded. ALPM-2 assumed that tough-mindedness corresponded to a disregard for others' contributions. In contrast, the results of the experiment suggest that it is *perceived* to be something akin to argumentativeness, or 'contrariness'.

In the present experimental design, opinion polarities were assigned randomly. For future work, it would be interesting to examine whether controlling for agreement versus disagreement of tough-minded characters might make a difference, i.e., whether a character that stuck to their interlocutor's topic and always agreed would also be considered tough-minded.

The results were stable across the native, non-native and all speakers conditions, with reduced significance levels for the native and non-native P trials, indicating that the perception of personality effects in this study was possible for native and non-native speakers alike.

An obvious limitation of ALPM-2 is that it relies on the existence of utterances to be ranked and chosen as the output. It would be much more interesting to be able to generate a number of paraphrases from a semantic representation, and choose amongst them. Thus, the following chapter explores ALPM-3, a deeper approach to modelling personality, in which utterances will be generated from scratch.

Chapter 4

Generating Text That Projects Personality

4.1 Summary

Chapter 3 was concerned with the ranking of arbitrary given utterances by personality using a set of linguistic and psycholinguistic features. This chapter defines and explores the Affective Language Production Model version 3 (ALPM-3), a significantly more flexible approach that generates a set of paraphrases from a semantic representation and ranks these paraphrases using n-gram language models to choose an utterance that best matches personality parameter settings. Utterances are then once more combined into dialogues between computer characters. In a web experiment similar to the one described in the previous chapter, generated dialogues were presented to human judges in order to assess whether they could perceive personality differences. It was found that native speakers were able to detect personality tendencies according to the model's expectations. Agreeableness was identified most accurately, while conscientiousness was the most difficult dimension to detect. A computer character's strategy to stick to their own topic agenda was perceived as Low A and High C. The dialogues were judged to be reasonably naturalistic.

The utterance ranking approach discussed in this chapter was developed by Carsten Brockmann, Amy Isard, Jon Oberlander and Michael White, implemented by Amy Isard with contributions from Carsten Brockmann and evaluated by Carsten Brockmann. The approach was described in Isard et al. (2006).

4.2 Introduction

Statistical approaches to natural language generation (NLG) have explored overgenerate-and-rank techniques for surface realisation. For example, the NITROGEN generator (Langkilde and Knight 1998) accepts an underspecified semantic representation and produces a word lattice of possible renderings that express the desired meaning. A statistical extractor then chooses the most likely candidate according to a language model. In instance-based NLG (Varges and Mellish 2001), candidates are ranked by comparing them to a database of stored instances.

This chapter investigates by way of the Affective Language Production Model version 3 (ALPM-3) how the overgenerate-and-rank approach can be employed to model personality in NLG. The OPENNLP CCG Library (OPENCCG, White 2006b) offers a suitable parsing and generation framework for the Combinatory Categorial Grammar (CCG, Steedman 2000) grammar formalism. Its realiser component takes a logical form as input and outputs a list of candidate sentences ranked using one or more n-gram language models. ALPM-3 proposes a combination of n-gram language models to choose the best utterance according to a character's personality and agenda.

OPENCCG supports *mild* overgeneration, which means that overgeneration can be restricted to certain parts of the grammar which are difficult to capture in rules and which benefit from n-gram ranking, e.g., adjectival and adverbial modifiers. It also employs an *anytime* search method: The generation process can be stopped at any time and the best edges found so far will be ranked highest in the then current agenda.

As a baseline, ALPM-2 assigned equal weight to all features. The n-gram language models in ALPM-3 can be seen as a more principled way of assigning appropriate weight to personality-related language differences.

4.2.1 Hypothesis

We hypothesise that personality can be recognisably projected through ALPM-3, an overgenerate-and-rank approach to realisation as part of an NLG system. The model is defined and evaluated in the following sections.

Topic	Nouns to Realize Topic
ACTION SCENES	*action scenes*
CHARACTERS	*characterization*
DIALOGUE	*dialogue*, *script*
FILM	*film*, *movie*
MUSIC	*music*, *score*
PLOT	*plot*, *story*
SPECIAL EFFECTS	*special effects*

Table 4.1: Topics available in the CrAg OpenCCG grammar, with nouns used to realise them.

4.3 Methods

4.3.1 OpenCCG Grammar

In order to generate utterances for characters discussing movies, a grammar from the COnversational Multimodal Interactions with Computers project (COMIC, Foster and White 2005) that is distributed with OpenCCG was adapted to the movie review domain. The list of topics that characters can discuss is specified in Table 4.1, along with nouns that can be chosen to realise these topics. It is similar to the one used for ALPM-1 and ALPM-2 (see Section 3.3.1.2.1). In this case, however, the topics are encoded as semantic classes. When realising a sentence, the semantic class can be specified in the logical form, and paraphrases are generated with all nouns that fit the semantic class.

The grammar was also extended with verbs appropriate for movie review discussions. Positive sentiments can be expressed with *to adore*, *to enjoy*, *to like* or *to love*; negative sentiments with *to dislike*, *to hate* or *to mind*. Also, *to think* is available as a propositional attitude verb which takes a sentence as a complement. A number of positive and negative adjectives were added.

Furthermore, there are adverbs, e.g., *actually*, *really* or *totally*, general multiword expressions, e.g., *I mean*, *kind of*, or *to be honest*, and canned full utterances, either general, e.g., *It was a bit dull.*, or movie-specific, e.g., *Some of the fight scenes were computer generated images of this guy in the suit*.

4.3.2 N-Gram Language Models

The basic assumption underlying ALPM-3 is that linguistic personality can be modelled by the combination of a variety of n-gram language models.

Language models are trained on a corpus and subsequently used to compute probability scores of word sequences. An n-gram language model approximates the probability of a word given its history of the preceding $n-1$ words. According to the chain rule, probabilities are then combined by multiplication. Equation (4.1) shows a trigram model that takes into account two words of context to predict the probability of a word sequence w_1^n:

$$P(w_1^n) \approx \prod_{i=1}^{n} P(w_i|w_{i-2}^{i-1}) \tag{4.1}$$

4.3.2.1 Avoiding the Length Effect

Because word probabilities are always less than 1 and therefore each multiplication decreases the total, if we use the standard model, longer sentences will always receive lower scores. This is known as the *length effect*. We therefore normalise by calculating the probability of a sentence as the geometric mean (GM) of the probability of each word in the sentence as shown in Equation (4.2):

$$P(w_1^n) \approx \prod_{i=1}^{n} P(w_i|w_{i-2}^{i-1})^{\frac{1}{n}} \tag{4.2}$$

4.3.2.2 Linear Combination of Language Models

OpenCCG supports the linear combination of language models, where each model is assigned a weight. For uniform interpolation of two language models P_a and P_b, each receives equal weight:

$$P(w_i|w_{i-2}^{i-1}) = \frac{P_a(w_i|w_{i-2}^{i-1}) + P_b(w_i|w_{i-2}^{i-1})}{2} \tag{4.3}$$

In the more general case, the language models are assigned weights λ_i, the sum of which has to be 1:

$$P(w_i|w_{i-2}^{i-1}) = \lambda_1 P_a(w_i|w_{i-2}^{i-1}) + \lambda_2 P_b(w_i|w_{i-2}^{i-1}) \tag{4.4}$$

For example, setting $\lambda_1 = 0.9$ and $\lambda_2 = 0.1$ assigns a high weight to the first language model.

4.3.2.3 OPENCCG N-Gram Ranking

In the OPENCCG framework, language models can be used to influence the chart-based realisation process. The agenda of edges is re-sorted according to the score an edge receives with respect to a language model. For ALPM-3, many paraphrases are generated from a given logical form, and they are then ranked in order of probability according to the combination of n-gram models appropriate for the character and stage of the dialogue.

The SRI Language Modeling Toolkit (SRILM, Stolcke 2002) is used to compute language models. All models for ALPM-3 are trigram models with backoff to bigrams and unigrams.

4.3.3 Personality Models

We experimented with two strategies for creating personality models. The first strategy involves using typical language for each personality dimension, and the second uses the language of one individual.

4.3.3.1 Building a Five-Factor Model

Nowson (2006) performed a study on language use in weblogs. The weblog authors were asked to complete personality questionnaires based on the five-factor model (see Section 2.2.1.2). All weblog authors scored high or medium on the openness dimension, so there were no data for typical low O language.

The data were divided into high, medium and low bands for each personality dimension. Language models were trained on each band so that the probability of a word sequence given a personality type could be assessed. The data from each individual weblog were used 5 times, once for each dimension.

For each personality dimension, the system simplifies a character's personality setting x by assigning a value of low ($x \leq 30$), medium ($30 < x \leq 70$) or high ($x > 70$). The five models corresponding to the character's assigned personality are uniformly interpolated to give the final personality model. If the character has been given a low O score, since there is no model for this personality type, only the other four models are interpolated.

4.3.3.2 Using an Individual's Language

An alternative strategy was to train n-gram models on the language of the individuals from the CrAg Corpus (see Section 3.3.1) and to use one of these models for each character in the dialogue. Table 4.2 shows the NEO-PI-R personality questionnaire results for the 20 participants that contributed to the CrAg Corpus. The participants' names are anonymised with random combinations of two letters.

4.3.4 Base Language Model

ALPM-3 incorporates a base language model to fall back to for probabilities unbiased by personality and to provide scores for words which might not exist in the personality-specific models. The five-factor personality language models use the uniform interpolation of a general language model (LM) based on data from the SWITCHBOARD corpus (Stolcke et al. 2000) and an LM built from the CrAg Corpus. For individual language models, the base model is calculated from the SWITCHBOARD corpus alone because the individual's language is already taken from the CrAg Corpus.

4.3.5 Character Specification and Dialogue Generation

Dialogues are composed by the Critical Agent Dialogue system version 2 (CRAG 2), a Java program that provides the framework for generating dialogues between two computer characters discussing a movie. The characters are parameterised for their personality by specifying values (on a scale from 0 to 100) for the five dimensions: E, N, A, C and O. Also, each character receives an agenda of topics they wish to discuss, along with polarities (POSITIVE/NEGATIVE) that indicate their opinion on the respective topic.

The character with the higher E score begins the dialogue, and their first topic is selected. Once an utterance has been generated, the other character is selected, and the system applies the algorithm shown in (4.5) to decide which topic should come next. This process continues until there are no topics left on the agenda of the current speaker.

(4.5) **if** (A $<$ 46) or (C $<$ 46) or (number of utterances about this topic $=$ 2)
then take next topic from own agenda
else continue on same topic

Person	E	N	A	C	O	Extremes
bb	58	*37*	60	41	67	Low N
bc	55	45	54	**57**	70	High C
bq	58	45	**65**	**57**	70	High A, High C
ci	44	56	62	52	72	
dd	46	53	55	41	64	
ds	46	56	42	36	62	
dt	51	43	57	41	65	
gq	**63**	49	54	26	67	**High E**
io	57	47	55	48	70	
ji	*26*	**68**	55	*21*	57	Low E, High N, Low C
lu	43	62	48	48	**75**	High O
mn	*31*	65	55	25	*53*	*Low E*, Low O
nd	51	60	55	**57**	70	High C
nx	53	48	57	46	65	
sd	46	**73**	*34*	*14*	*55*	**High N**, Low A, Low C, Low O
si	34	45	59	**67**	57	High C
ss	43	*40*	*36*	24	70	*Low N*, Low A
st	**62**	44	53	50	67	High E
ud	36	48	64	31	**74**	High O
uq	48	47	**67**	46	**74**	High A, High O

Table 4.2: NEO-PI-R scores for the persons in the CrAg Corpus. Extreme personalities are emphasised: In each personality dimension column, the lowest two values are italicised and the highest two values are set in bold face. Emphasis in the rightmost column indicates which individuals' language models were chosen to contrast high and low E and N.

```
<utterance>
  <utt topic="music" polarity="dislike" opp-polarity="like"
       so="no" right="no" also="no" well="yes" and="no"
       but="no">
    <pred adj="bad" />
    <opp-pred adj="good" />
  </utt>
</utterance>
```

Figure 4.1: Simple utterance specification.

The system creates a simple XML representation of the character's utterance, using the specified topic and polarity. An example using the topic MUSIC and polarity NEGATIVE is shown in Figure 4.1. At this point the system also decides which discourse connectives may be appropriate, based on the previous topic and polarity.

4.3.5.1 OPENCCG Logical Forms

Following the method described in Foster and White (2004), the basic utterance specification is transformed, using stylesheets written in the Extensible Stylesheet Language Transformations (XSLT) language, into an OPENCCG logical form. We make use of the facility for defining optional and alternative inputs (White 2006a) and underspecified semantics to mildly overgenerate candidate utterances. A fragment of the logical form which results from the transformation of Figure 4.1 is shown in Figure 4.2.

Optional interjections (*I mean*, *you know*, *sort of*) and conversational markers (*right*, *but*, *and*, *well*) are added where appropriate given the discourse history. When the full logical form is processed by the OPENCCG system, the output consists of sentences of the types shown in (4.6):

(4.6) (I think) the music was bad.
(I think) the music was not (wasn't) good.
I did not (didn't) like the music.
I hated the music.
One thing I did not (didn't) like was the music.
One thing I hated was the music.

```
<node id="l1:opinion" pred="like" tense="past">
  <rel name="Speaker">
    <node id="p1:person" pred="pro1" num="sg" />
  </rel>
  <rel name="Content">
    <node id="f1:cragtopic" pred="music" det="the" num="sg" />
  </rel>
  <opt>
    <rel name="Modifier">
      <node id="w1:adv" pred="well" />
    </rel>
  <opt>
  <opt>
    <rel name="HasProp">
      <node id="a2:proposition" pred="kind-of" />
    </rel>
  </opt>
  <opt>
    <rel name="Modifier">
      <node id="a1:adv" pred="you-know" />
    </rel>
  </opt>
</node>
```

Figure 4.2: Fragment of the logical form generated from the utterance specification in Figure 4.1.

The fragmentary logical form in Figure 4.2 would create all possible paraphrases from (4.7):

(4.7) (well) (you know) I (kind of) {liked, loved} the {music, score}

By using synonyms (e.g., *plot* = *story*, *comedy* = *humour*) and combining the sentence types and optional expressions, up to 3000 possibilities are created per utterance, and the best candidate is chosen by the specific combination of n-gram models appropriate for the given personality and dialogue history, as described in Sections 4.3.2 to 4.3.4.

Topic	Polarity
FILM	NEGATIVE
ACTION SCENES	POSITIVE
PLOT	NEGATIVE
DIALOGUE	POSITIVE
SPECIAL EFFECTS	NEGATIVE

Table 4.3: Topics and polarities for the pretest to determine the personality model weight that maximises variety.

4.3.6 Materials

4.3.6.1 Choosing a Weight for the Personality Model

In ALPM-3, a base language model is combined with a personality language model. A pretest was conducted in order to determine how the weights given to the base and personality models influence the variety of the generated language, and in order to choose the weight that maximises this variety.

Five topics and polarities were randomly chosen (see Table 4.3). Utterances were generated for high E and high N characters, with the weight assigned to the personality model varying from 0 to 1 in steps of 0.1 and, accordingly, the weight for the base language model being decreased from 1 to 0.

To evaluate the variety of generated utterances, we looked at the top 100 edges produced by the OPENCCG realiser. For this study, the ranking produced with a weight of 0 for the personality model (and 1 for the base language model) was taken as the baseline. This was compared to the ranking produced with each of the increasing weight settings for the personality model by way of Spearman's rank correlation ρ.

The utterances in both rankings were numbered from 1 to 100. Utterances that occurred in only one of the rankings were appended to the respective other ranking in their original order and received numbers higher than 100. From these two rankings, a combined list was produced which also included information about the utterances' rank differences. If, for example, an utterance had rank 20 in the first list and rank 60 in the second list, its rank difference would be $60 - 20 = 40$. Spearman's ρ was computed from the two rankings and served as a measure of how much the rankings

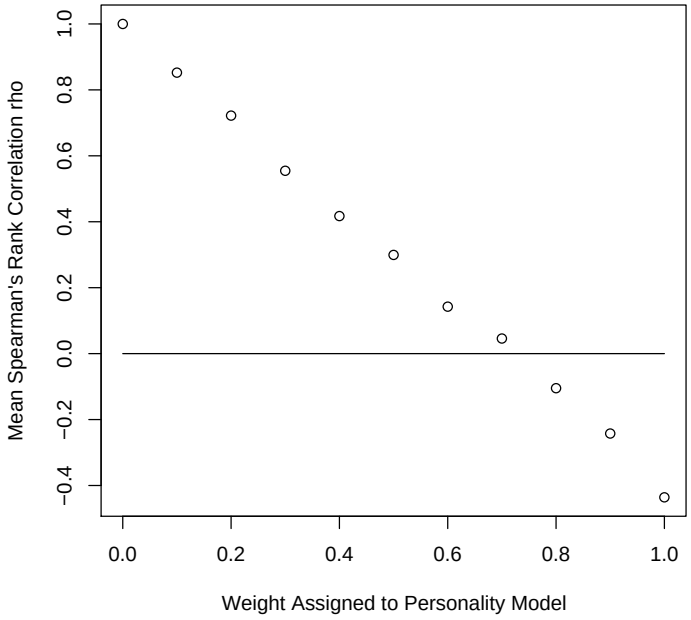

Figure 4.3: Mean Spearman's rank correlation ρ at varying personality model weight settings: Five-factor model and E dimension.

differed from each other.

Five-factor personality models were built from the weblog data for a high E character (E = 75, N = 50, A = 50, C = 50, O = 50) and a high N character (E = 50, N = 75, A = 50, C = 50, O = 50). Figures 4.3 and 4.4 show the mean rank correlation values in relation to personality model weight settings for the high E and high N characters, respectively.

For generation with language models based on the language of individuals from the CrAg Corpus, participant *gq* was chosen for high extraversion and participant *sd* was chosen for high neuroticism. Figures 4.5 and 4.6 show the mean rank correlation values in relation to personality model weight settings for the high E and high N characters, respectively.

A positive correlation indicates that utterances were ordered similarly, while a negative correlation suggests that they were ordered in reverse order. A correlation close to zero means that the orderings are unrelated and that the second ranking is independent from the baseline.

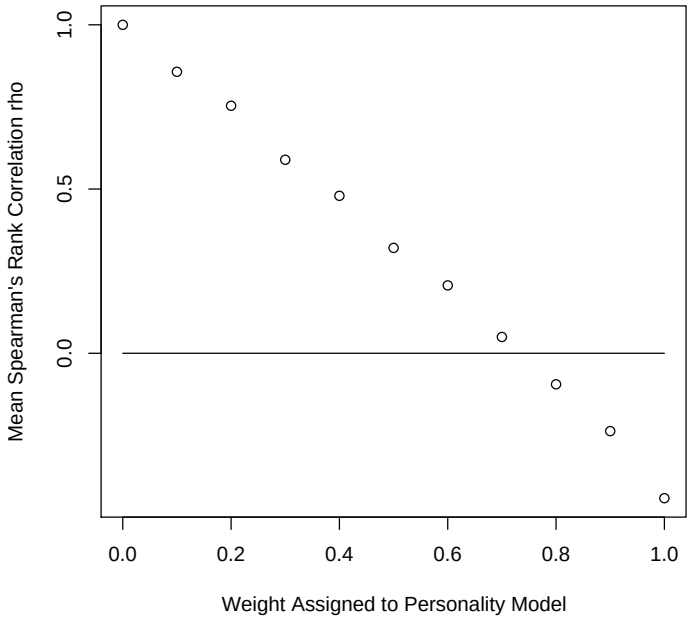

Figure 4.4: Mean Spearman's rank correlation ρ at varying personality model weight settings: Five-factor model and N dimension.

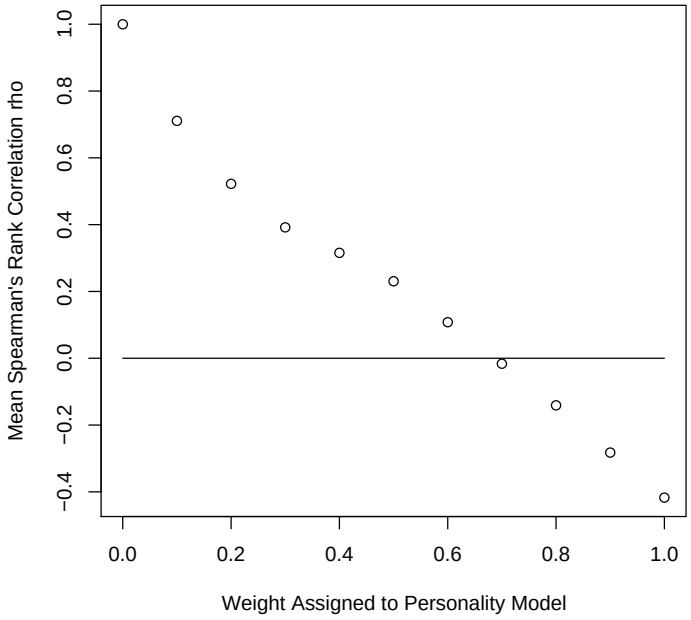

Figure 4.5: Mean Spearman's rank correlation ρ at varying personality model weight settings: CrAg Corpus individual *gq*'s language model and E dimension.

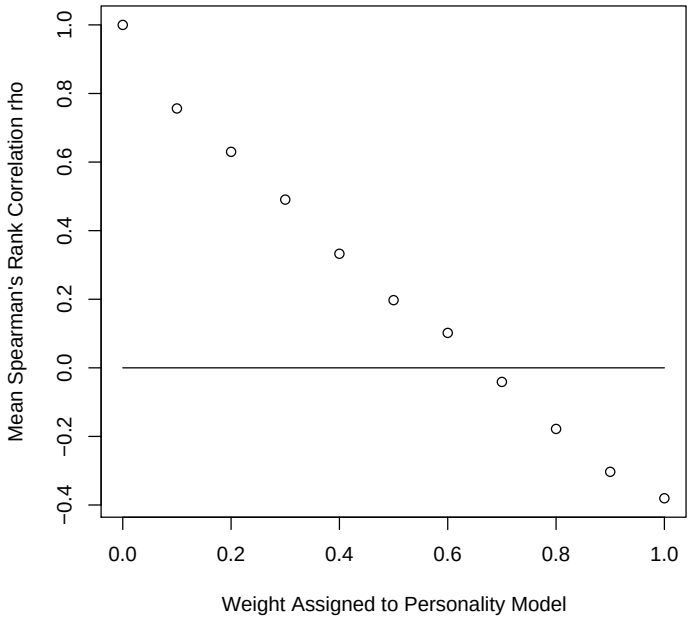

Figure 4.6: Mean Spearman's rank correlation ρ at varying personality model weight settings: CrAg Corpus individual *sd*'s language model and N dimension.

Utt.	Five-Factor		Individual	
No.	E	N	E	N
1	-0.2669	-0.2533	-0.3868	-0.4558
2	-0.5000	-0.5000	-0.5000	-0.5000
3	-0.2734	-0.2720	-0.3536	-0.4233
4	-0.3943	-0.3935	-0.4602	-0.5000
5	-0.5000	-0.5000	-0.4946	-0.5000
$\overline{x}$	-0.3869	-0.3838	-0.4391	-0.4758
s	0.1150	0.1190	0.0657	0.0351

Table 4.4: Spearman's rank correlation ρ for the test utterances generated with high versus low E and N settings, for both five-factor and individual personality language models.

As can be seen from the figures, across all four conditions, a personality model weight of 0.7 leads to a Spearman's rank correlation close to 0, which means that the utterance ordering is maximally independent from the one created by the base language model. For the generation of dialogues to be presented to human judges, we are interested in making personality as recognisable as possible. In order to achieve maximum impact, we therefore set the personality model weight to 0.7 in all further experiments.

4.3.6.2 Choosing a Personality Model

A second pretest was performed in order to find out which of the two personality models would be more effective in capturing differences between extreme personality settings. The weight for each personality model was fixed at 0.7, in accordance with the results from the previous section.

This time, the five test utterances (see Table 4.3) were generated in four conditions: Extraversion and neuroticism for both five-factor and individual personality language models. In each condition, high and low extremes were contrasted. For the five-factor models, E and N were set to 75 and 25, respectively, with the other dimensions constant at 50. For the models based on individuals, *gq* and *mn* were contrasted for E, and *sd*

and *ss* were contrasted for N. Spearman's rank correlation ρ was computed to assess the amount of variation. Table 4.4 shows the results.

The lower the correlation, the more the respective model changes and re-orders candidate utterances between high and low E or N language. The minimum possible correlation is -0.5 because utterances occurring in only one of the lists are appended to the other list before computing the rank correlation, and if there is no overlap between the initial lists, the top 100 utterances of the first list will appear as the second 100 utterances of the other list and vice versa.

The correlations for the individual language models were consistently lower or at least equal to the five-factor models, apart from case 5 in the E dimension. For E, the mean ρ value in the individual condition ($\bar{x} = -0.4391$, $s = 0.0657$) was lower than the mean for the five-factor condition ($\bar{x} = -0.3869$, $s = 0.1150$). However, the decrease was not significant, $t(4) = 2.1646$, $p = 0.0964$, related, two-tailed. For N, the mean ρ value in the individual condition ($\bar{x} = -0.4758$, $s = 0.0351$) was once again lower than the mean for the five-factor condition ($\bar{x} = -0.3838$, $s = 0.1190$). The decrease was also not significant, $t(4) = 2.2712$, $p = 0.0856$, related, two-tailed.

The five-factor models were chosen to be used in further experiments. The lack of significance between the mean rank correlations indicates that both approaches behave similarly. Also, individual language models are more contaminated than the five-factor ones because they vary on multiple dimensions and cannot be controlled as precisely. Finally, the individual models are based on less data than the five-factor models.

4.3.6.3 Generation

Dialogues were generated in four different conditions, as shown in Table 4.5. Each condition sets the two computer characters to opposing extremes on either the E or the N dimension, while keeping the respective other dimension at a middle level. Furthermore, character A is always Low A and C, and character B is always High A and C. All characters are set to Mid O.

This experiment is designed in parallel to the ALPM-2 evaluation described in Section 3.3.4. The difference is that the present experiment is based on the five-factor model of personality. We assume that high psychoticism as in the three-factor model can be approximated by low agreeableness and conscientiousness, and vice versa, following previous studies (Gill 2004, p. 16):

> The simplest interpretation of these two models maps the NEO-PI-R traits conscientiousness and agreeableness negatively onto EPQ-R Psychoticism.

		Personality Parameter Setting				
Condition	Character	E	N	A	C	O
A) High E	A	75	50	25	25	50
vs. Low E	B	25	50	75	75	50
B) Low E	A	25	50	25	25	50
vs. High E	B	75	50	75	75	50
C) High N	A	50	75	25	25	50
vs. Low N	B	50	25	75	75	50
D) Low N	A	50	25	25	25	50
vs. High N	B	50	75	75	75	50

Table 4.5: Personality parameter settings for the four experimental conditions.

Two dialogues were generated per condition. The movie under discussion and the characters' agenda and their opinions about the topics were randomly assigned. Each dialogue was eight utterances long, with characters taking turns, each of them producing four utterances altogether.

In order to maximise the variation of the generated language, the system does not simply choose the highest-ranked utterance for a given personality parameter configuration. Instead, if, e.g., a High E utterance is required, lists of 100 candidates of both High and Low E versions are generated. These lists are then sorted by their rank differences as in the preparation for the rank correlation calculation for the weight-choosing pretests (see Section 4.3.6.1). The utterance finally chosen is the one at the bottom or top of this rank difference ordered list, depending on whether a High or Low personality version is required.

4.3.6.4 Web Experiment

The experiment was conducted on the World Wide Web, advertised via the Language Experiments Portal[1]. An introductory page explained the task and defined the four

[1] http://www.language-experiments.org/

personality dimensions of interest (E, N, A and C) by paraphrasing descriptions from Buchanan et al. (1999):

Extraversion This trait reflects preference for, and behavior in, social situations. People high in extraversion are energetic and seek out the company of others. Low scorers (introverts) tend to be more quiet and reserved.

Emotional instability This trait reflects the tendency to experience negative thoughts and feelings. High scorers are prone to insecurity and emotional distress. Low scorers tend to be more relaxed, less emotional and less prone to distress.

Agreeableness This trait reflects how we tend to interact with others. People high in agreeableness tend to be trusting, friendly and cooperative. Low scorers tend to be more aggressive and less cooperative.

Conscientiousness This trait reflects how organised and persistent we are in pursuing our goals. High scorers are methodical, well organised and dutiful. Low scorers are less careful, less focussed and more likely to be distracted from tasks.

The introductory page also contained an example dialogue similar to those that would be presented during the experiment. The complete instructions are reproduced in Section A.2 in the appendix.

Subsequently, the participants were presented with one dialogue per condition, randomly chosen from the pool of available dialogues, in random order.[2] One dialogue was shown at a time, and the participants were asked to judge which of the two characters scored higher on the E, N, A and C dimensions. The higher-scoring character had to be chosen; *equal* was not an option. The judges also assessed, on a seven-point Likert scale, how well the characters got on with each other and how smoothly the conversation went, with one point representing *very badly* or *not at all smoothly* and seven points representing *very well* or *very smoothly*, respectively. The participants were asked to rate each dialogue independently from the others.

Getting on and *smoothness* were introduced purely for *internal* comparison. As in ALPM-2, generated dialogues lack contextuality and are therefore definitely not going to be as smooth as real dialogue, such as that found in the CrAg Corpus.

Five exemplars for each personality dimension were displayed at the bottom of the page (see Table 4.6). The adjectives were taken from a study by Goldberg (1992,

[2]Note that participants were also presented with another set of generated dialogues, manipulated to reflect the phenomenon of alignment as discussed in Chapter 6. Results on these materials are presented there.

Dimension	Exemplars
Extraverted	Talkative, bold, assertive, unrestrained, verbal.
Emotionally unstable	Moody, jealous, temperamental, touchy, high-strung.
Agreeable	Kind, sympathetic, warm, pleasant, helpful.
Conscientious	Organised, efficient, neat, systematic, thorough.

Table 4.6: Exemplars for the personality dimensions displayed during the web experiment, taken from Goldberg (1992, p. 34 f.).

p. 34 f.). For each dimension, the five adjectives with the highest varimax-rotated factor loadings in descriptions of liked peers were selected.

The experiment was open to native as well as non-native speakers of English to allow an analysis of whether the results held for both groups. Participants were entered into a draw for a £15 Amazon.co.uk gift certificate if they chose to specify a valid e-mail address. All data were analysed anonymously.

4.3.6.5 Example Dialogues

To give an impression of what the generated dialogues look like, this section presents one dialogue each of the A (High E versus Low E) and C (High N versus Low N) conditions. The parameter settings were not shown to the experiment's participants. Below, the polarities POSITIVE and NEGATIVE are abbreviated as $+$ and $-$, respectively.

Note that unlike with ALPM-2, extraversion is no longer easily recognisable from the length of the utterances. This is a consequence of the language modelling approach that is hard to avoid; even though scores are normalised by using the geometric mean (see Section 4.3.2.1), short utterances are still preferred in general. The Low A/Low C character's strategy of switching topics in order to follow his own agenda is obviously recognisable, as it was in the previous study for High psychoticism.

The LMs' influence on language choice can be seen in the first example dialogue below; in three of his four utterances, the High E character A uses the phrase *I mean I didn't think*, while the Low E character B prefers *I (kind of) thought* throughout. In the second example dialogue, there is an influence on lexical choice: The High N character A refers to the film as *film*, while the Low N character B uses *movie*.

4.3.6.5.1 High E versus Low E

Parameter Settings

Character A's personality High E, Mid N, Low A, Low C, Mid O.

Character A's agenda PLOT/−, CHARACTERS/−, MUSIC/−, FILM/−.

Character A's further opinions SPECIAL EFFECTS/−, ACTION SCENES/+, DIALOGUE/−.

Character B's personality Low E, Mid N, High A, High C, Mid O.

Character B's agenda ACTION SCENES/+, SPECIAL EFFECTS/+, PLOT/−, DIALOGUE/−.

Character B's further opinions CHARACTERS/−, FILM/−, MUSIC/−.

Movie Mystic River.

Base language model SWITCHBOARD LM, weight 0.5; CrAg Corpus LM, weight 0.5.

Combined base and personality language model Base LM, weight 0.3; personality LM, weight 0.7.

Dialogue

A (PLOT) "I mean I didn't think that the story was great."

B (PLOT) "Yes I mean I thought that it wasn't nicely tied up."

A (CHARACTERS) "And I mean I thought there was no characterization at all."

B (CHARACTERS) "Yeah there was no characterization at all I mean."

A (MUSIC) "I mean I didn't think that the music was great."

B (MUSIC) "Yeah I kind of thought that the music was also very bad you know."

A (FILM) "I mean I didn't think the film was good."

B (FILM) "I mean I kind of thought that it was a bit dull."

4.3.6.5.2 High N versus Low N

Parameter Settings

Character A's personality Mid E, High N, Low A, Low C, Mid O.

Character A's agenda DIALOGUE/−, FILM/−, PLOT/−, ACTION SCENES/+.

Character A's further opinions MUSIC/−, CHARACTERS/+, SPECIAL EFFECTS/+.

Character B's personality Mid E, Low N, High A, High C, Mid O.

Character B's agenda ACTION SCENES/+, PLOT/+, SPECIAL EFFECTS/−, CHARACTERS/+.

Character B's further opinions MUSIC/+, DIALOGUE/+, FILM/+.

Movie Intolerable Cruelty.

Base language model SWITCHBOARD LM, weight 0.5; CrAg Corpus LM, weight 0.5.

Combined base and personality language model Base LM, weight 0.3; personality LM, weight 0.7.

Dialogue

A (DIALOGUE) "So I didn't think that the script was very good you know."

B (DIALOGUE) "Well actually to be honest I didn't think that the script was very bad."

A (FILM) "I mean I didn't think the film was very good."

B (FILM) "Well actually I mean I didn't think that the movie was very bad."

A (PLOT) "But you know I didn't think that the story was very good."

B (PLOT) "I mean I thought the story was really good."

A (ACTION SCENES) "I mean I didn't think that the action scenes were very bad really."

B (ACTION SCENES) "Yeah I kind of thought that the action scenes were very good also you know."

4.4 Results

Eighty participants completed the web experiment. The data were filtered to exclude replies that were submitted after less than five minutes (five cases) or more than 45 minutes (one case), leaving 74 judgements for further analysis. Of these, 39 were native and 35 were non-native speakers of English.

The data were evaluated by comparing the settings used to generate the experimental materials with the actual human judgements. The hypothesis was that for conditions A and B people would be able to detect differences in extraversion, while for conditions C and D they would see differences in neuroticism. Differences in agreeableness and conscientiousness were expected across all four conditions. Table 4.7 shows the results of this comparison. Figures 4.7 to 4.9 show the percentage of participants' choices in agreement with expectations for the native, non-native and all speakers conditions, respectively.

	Native Speakers				Non-Native Speakers				All			
Choice	E	N	A	C	E	N	A	C	E	N	A	C
Expected	43	46	107	82	33	41	98	55	76	87	205	137
Not expected	35	32	49	74	37	29	42	85	72	61	91	159
Number of trials	78	78	156	156	70	70	140	140	148	148	296	296
Expected (%)	55.1	59.0	68.6	52.6	47.1	58.6	70.0	39.3	51.4	58.8	69.3	46.3
Significance			***				***	*		*	***	

Table 4.7: Numbers of expected versus not expected choices for the web experiment. Significance levels of binomial test: $*$: $p \leq 0.05$, $**$: $p \leq 0.01$, $***$: $p \leq 0.001$.

Binomial tests were performed to assess whether the distribution of values differed significantly from chance. Native English speakers agreed with ALPM-3 more than 50% of the time across all four personality dimensions, in the order A > N > E > C, although significance was reached only for A ($p \leq 0.001$). Non-native speakers' judgements followed the same pattern, A > N > E > C, but the percentages were lower for E, N and C and higher for A. E and C agreement was less than 50%. Significance was reached for A ($p \leq 0.001$) and C ($p \leq 0.05$). All participants taken as a group agreed with ALPM-3 more than 50% of the time in the E, N and A conditions, with significance for A ($p \leq 0.001$) and N ($p \leq 0.05$).

Figures 4.10 to 4.12 show the distribution of ratings of how well the characters got on with each other and how smoothly the conversation went, for native English, non-native English and all speakers, respectively. The average native speakers' judgements for getting on (median $m = 5$, mean $\bar{x} = 4.404$, standard deviation $s = 1.422$) and smoothness ($m = 4$, $\bar{x} = 4.109$, $s = 1.672$) were slightly lower than the non-native judgements for getting on ($m = 4.5$, $\bar{x} = 4.493$, $s = 1.580$) and smoothness ($m = 5$, $\bar{x} = 4.593$, $s = 1.631$), respectively. The difference between means for getting on (0.089) was lower than for smoothness (0.484). The overall values for getting on ($m = 5$, $\bar{x} = 4.446$, $s = 1.497$) and smoothness ($m = 4$, $\bar{x} = 4.338$, $s = 1.667$) lay in between.

4.5 Discussion

For native speakers, the personality perception tendencies are in the expected direction across all four personality dimensions. This suggests that ALPM-3 models personality differences in generation appropriately. However, the effects were weak; the strength of the signal might not be sufficiently high. Non-native speakers recognised personality less well, which might indicate that they do not perceive subtle differences in the same way as native speakers.

Agreeableness seems to be the dimension that is the easiest to perceive, while conscientiousness is the most difficult. This indicates that our hypothesis of combining A and C to achieve the effect of psychoticism in the three-factor model needs revision. Relating these findings to the ALPM-3 topic selection strategy, people concur with us in assuming that agreeable speakers pick up the topic that their dialogue partner used in the previous utterance. However, contrary to our assumption, participants tend to consider it conscientious if speakers follow their own agenda, not caring about their dialogue partner.

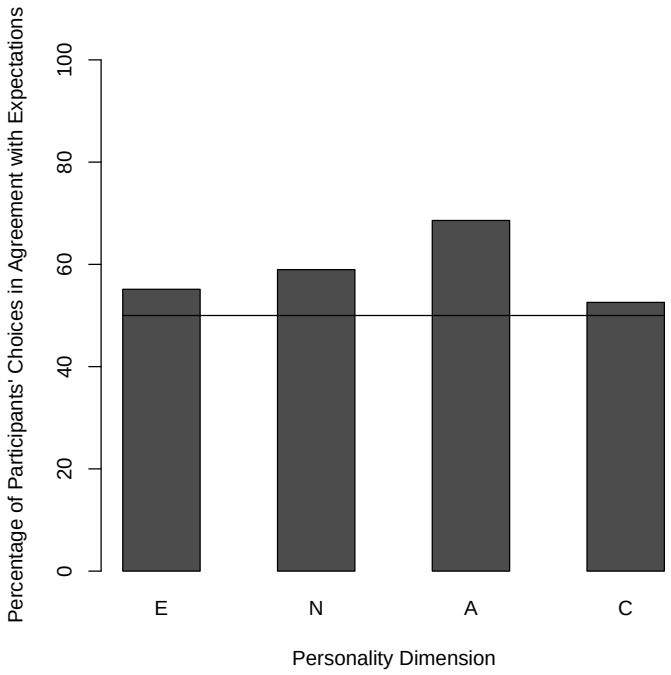

Figure 4.7: Percentage of native English speakers' choices in agreement with expectations, by personality dimension.

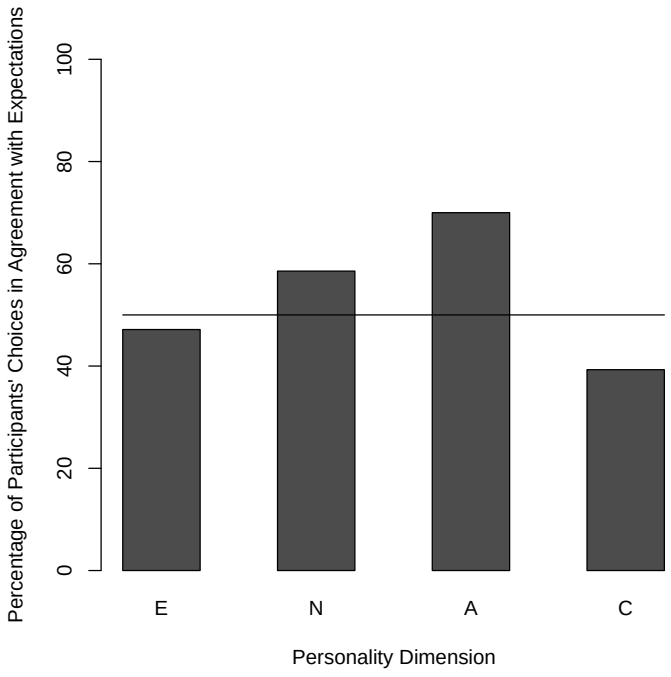

Figure 4.8: Percentage of non-native English speakers' choices in agreement with expectations, by personality dimension.

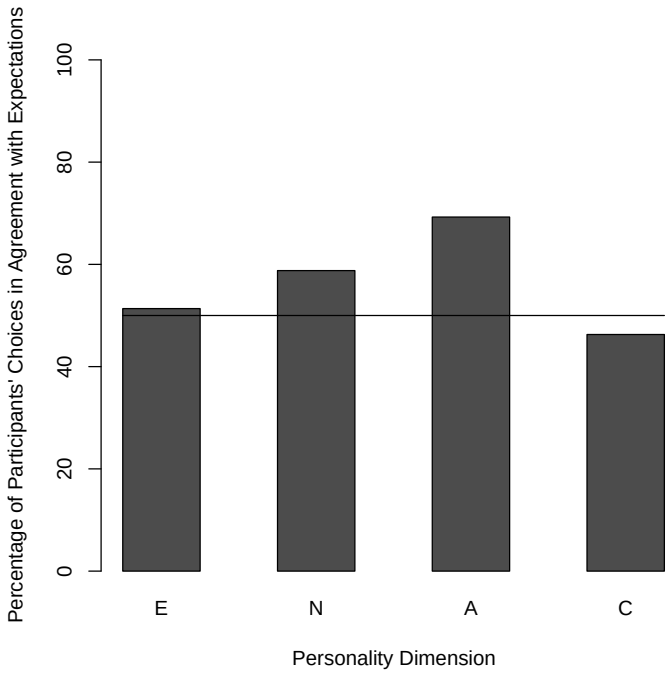

Figure 4.9: Percentage of all participants' choices in agreement with expectations, by personality dimension.

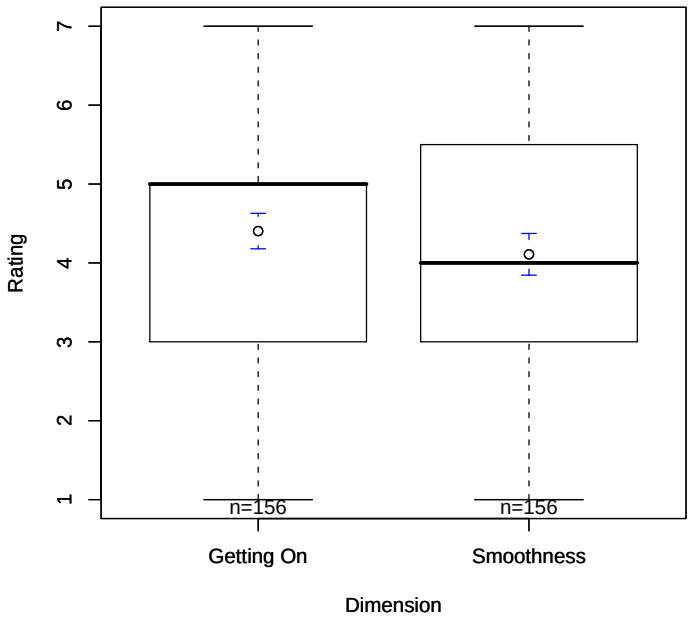

Figure 4.10: Native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went.

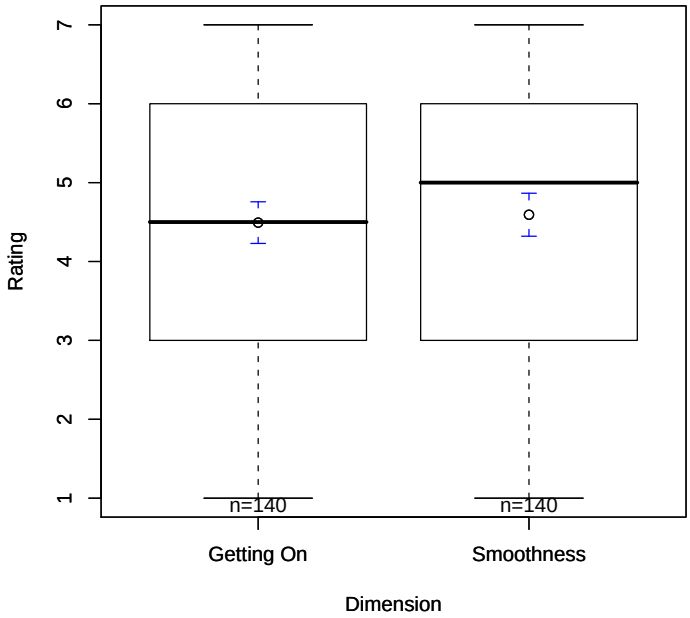

Figure 4.11: Non-native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went.

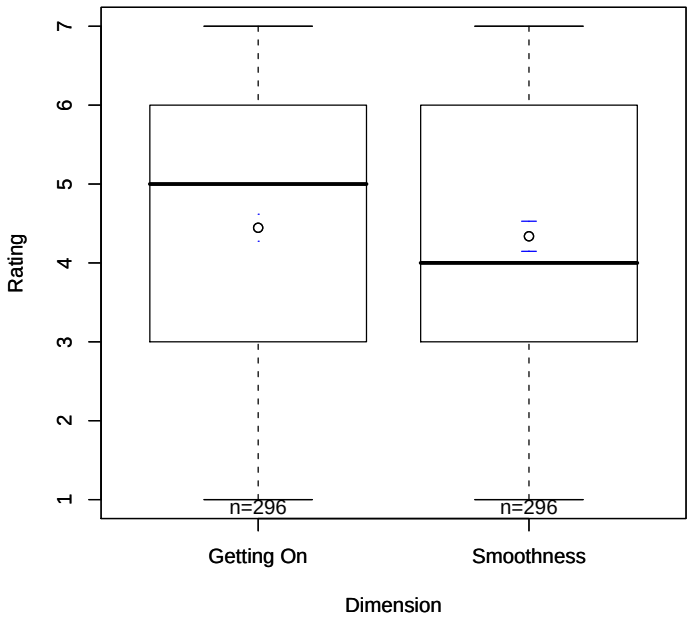

Figure 4.12: All participants' ratings of how well the characters got on with each other and how smoothly the conversation went.

There is evidence that the personality dimensions agreeableness, conscientiousness and openness are more important to success of long-term relationships than extraversion and neuroticism (Rammstedt and Schupp 2008). Hence, if building a computer agent companion, which uses language, it is worth getting its level of agreeableness right (and in particular, to reflect that of its human user) – especially if the companion is to be used over a long period of time.

Comparing the present results to the evaluation of ALPM-2, extraversion is no longer the easiest dimension to perceive; in fact, neuroticism was perceived more easily than extraversion. This might be due to the lack of variation in utterance length, as indicated in Section 4.3.6.5. Psychoticism in the previous experiment was perceived similarly to conscientiousness in the present study.

Given that the mean ratings of how well the characters got on with each other and how smoothly the conversation went were above the Likert scale's middle choice (four points), we have some confidence that the generation system produces output in which the dialogues are judged at least averagely on both of these dimensions, and that the output is reasonably naturalistic.

For future work, it would be interesting to explore whether it is possible to integrate ALPM-2 and ALPM-3 in order to draw on their respective strengths. The models' underlying implementations are very different; the simplest way to combine them might be a pipeline architecture where the top n results generated by the ALPM-3 module would be passed on to the ALPM-2 component for further scoring. However, the candidates produced by ALPM-3 would be very similar to each other due to OpenCCG's n-gram language model ranking, so ALPM-2 would probably have little influence on the outcome.

Pickering and Garrod (2004, 2006) argued that alignment is the basis of successful communication in dialogue. In a personality projection study, Gill et al. (2004) found that alignment behaviour in dialogue is related to a speaker's personality, in particular to the N dimension. Also, alignment might have an influence on the perception of our topic selection strategy. In the following chapter, we therefore introduce alignment and build a model to approximate results of psycholinguistic research. In Chapter 6, alignment is then incorporated into ALPM, and the new model is evaluated in comparison to the findings of the present chapter.

Chapter 5

Modelling Alignment with Cache Language Models

5.1 Summary

In the previous two chapters, we modelled personality-related individual *differences* in language behaviour. We now propose cache language models (CLMs) in order to find utterances that *match* at different linguistic levels. A CLM is trained on a prime utterance and then applied to rate target utterances.

To evaluate the approach, we examine the extent to which psycholinguistic findings on alignment and priming behaviour can be modelled computationally. Cleland and Pickering (2003) conducted experiments on how priming was influenced by the repetition of nouns and adjectives and by semantic relatedness of the nouns involved. Pickering and Branigan (1998) studied the priming of prepositional object/double object constructions. Materials from these experiments are re-used here to evaluate the cache language modelling technique. Results are presented, discussed and compared to the psycholinguistic findings, and limitations of the approach are pointed out. We also determine a weighting that maximises the matching effect and will be used in the experiment described in the next chapter.

The cache language model component discussed in this chapter was designed by Carsten Brockmann in collaboration with Michael White. It was initially implemented in Perl by Carsten Brockmann and later re-implemented in Java by Amy Isard with contributions from Carsten Brockmann. The approach was evaluated by Carsten Brockmann and described in Brockmann et al. (2005).

5.2 Introduction

According to the Interactive Alignment Model of dialogue processing (IAM, Garrod and Pickering 2004, Pickering and Garrod 2004, 2006), dialogue participants align their internal representations at different linguistic levels: Phonological, syntactic and semantic representations as well as their situation models. Alignment at lower levels leads to more alignment at higher levels. Alignment is an automatic process, and it ensures that dialogue partners operate on common representations.

On the one hand, there is a parity of representations that a person uses for language production and comprehension. On the other hand, there is priming of representations between speakers and listeners. These two principles lead to imitation and subsequently to alignment of representations between interlocutors.

With *lexical alignment*, interlocutors refer to particular objects or events using the same referring expressions, e.g., they agree on using the noun *movie* instead of *film* or the verb *to hand* instead of *to give*. *Syntactic alignment* occurs when speakers pick up their dialogue partner's syntactic structures.

Priming has been investigated in psycholinguistic studies. For example, Branigan et al. (2000) conducted an experiment where pairs of speakers described pictures to each other. One speaker was the experimenter's confederate; the syntactic structure of the confederate's description affected the other speaker's language choice. For example, if the confederate used a prepositional object construction like (5.1)(a) to describe his picture, the other speaker would be more likely to reply with a similar prepositional object sentence (5.1)(b) than to produce the double object version (5.1)(c).

(5.1) (a) The cowboy gave the banana to the burglar.

(b) The sailor gave the book to the teacher.

(c) The sailor gave the teacher the book.

In the course of this chapter, we will develop a computational model of a type of *matching* that approximates human performance in psycholinguistic priming experiments. The model will be integrated into the natural language generation system described in the previous chapter, and used in the following chapter to contrast personality-related language differences with alignment behaviour.

5.2.1 Hypothesis

We hypothesise that the effect of lexical and syntactic priming, and subsequently alignment, can be simulated with the cache language model matching approach.

5.3 Methods

5.3.1 Cache Language Models

The OPENCCG surface realiser (White 2006b) takes as input a logical form specifying the propositional meaning of a sentence, and returns a ranked list of surface strings that express this meaning according to the lexicon and grammar, where rankings are determined by n-gram language models derived from examples of desired realisations (reducing the need for hand-crafted rules). The n-gram models are employed in a best-first anytime search, in such a way that preferred realisations tend to be found early in the search process. It is possible to plug in n-gram models that interpolate a cache, with the effect that aligning realisations will be produced with less effort.

Cache language models work by interpolating simple language models derived from the recent context with more elaborate, context-independent models. The performance of language models can be compared according to the *perplexity* measure, i.e., the reciprocal of the average probability per word assigned to a sample text (Kuhn and De Mori 1990, p. 571). Lower perplexity indicates better performance. As Goodman (2001) explains, CLMs can yield impressive reductions in perplexity, and bigram and trigram CLMs usually work better than unigram ones.

For instance, Kuhn and De Mori (1990) developed a language model for speech recognition that combines a 3g-gram component with a cache component. The 3g-gram model consists of two parts: The first part assigns probabilities to part-of-speech (POS) categories and works just like a word-based trigram model, only with a vocabulary of POS tags; the second part computes probabilities of words based on their frequency within POS categories. This is interpolated with the cache model which keeps track of the 200 most recent words in each of 19 different POS categories. With this approach, higher probabilities are assigned to recently seen words, which results in a perplexity of less than a third of that of a pure 3g-gram model.

We use the SRI Language Modeling Toolkit (SRILM, Stolcke 2002) to compute n-gram language models and score test sentences. The toolkit includes a trigram language model derived from Switchboard (Godfrey et al. 1992) corpus data, which we

use as a default smoothed language model. To simulate alignment, we interpolate it with a language model calculated on the basis of a cached sentence. The cached sentence can be seen as the previous utterance in a dialogue.

In addition to standard CLM features, the psycholinguistic results which we want to emulate demand that the CLM support the following specific feature: Semantic relatedness. To that end, we annotate words with semantic classes (see below). Creating separate caches for different POS tags as in Kuhn and De Mori's model, however, is not required; we are only interested in the words and the corresponding semantic classes of the immediately preceding sentence.

5.3.2 Interpolating Cache and Default Language Models

By the chain rule, the probability of a word sequence $w_1, \ldots, w_n$ is equal to the product of the probabilities of each word w_i given the preceding ones, w_1^{i-1}:

$$P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1^{i-1}) \tag{5.2}$$

We approximate the probability of a word given its history by the probability of a word given the preceding two words, i.e., by using trigrams. In the trigram probability, we interpolate a CLM, $P_{\text{cache}}(w_i | w_{i-2}^{i-1})$, with the default smoothed model, $P_{\text{smooth}}(w_i | w_{i-2}^{i-1})$, as follows:

$$P(w_i | w_1^{i-1}) \approx \lambda P_{\text{cache}}(w_i | w_{i-2}^{i-1}) + (1 - \lambda) P_{\text{smooth}}(w_i | w_{i-2}^{i-1}) \tag{5.3}$$

The CLM gets weight λ, and the default model $1 - \lambda$. The CLM itself is the uniform interpolation of word- and class-based trigram models, as shown in (5.4) below.

$$P_{\text{cache}}(w_i | w_{i-2}^{i-1}) = \frac{P_{\text{word}}(w_i | w_{i-2}^{i-1}) + P_{\text{class}}(c_i | w_{i-2}^{i-1}) P(w_i | c_i)}{2} \tag{5.4}$$

Semantic classes are a way of grouping together lexical items with similar semantic properties; they provide a backoff mechanism if there is no exact word match. We assign simple classes to verbs, nouns and adjectives. Function words receive their own lexical entry as their semantic class; we do not put them in a separate group because we do not want them to prime each other. See Table 5.1 for some examples of the assignment of semantic classes.

The word- and class-based trigram models are themselves uniform interpolations

Lexical Item	Semantic Class
gave	EXCHANGE
handed	EXCHANGE
's (contracted form of *is*)	STATE
businessman	PERSON
secretary	PERSON
diamond	SHAPE
square	SHAPE
red	COLOUR
green	COLOUR
that	THAT
the	THE

Table 5.1: Example semantic classes assigned to lexical items.

of unigram, bigram and trigram probabilities:

$$P_{\text{word}}(w_i|w_{i-2}^{i-1}) = \frac{P(w_i) + P(w_i|w_{i-1}) + P(w_i|w_{i-2}^{i-1})}{3} \tag{5.5}$$

$$P_{\text{class}}(c_i|w_{i-2}^{i-1}) = \frac{P(c_i) + P_{\text{bi}}(c_i|w_{i-1}) + P_{\text{tri}}(c_i|w_{i-2}^{i-1})}{3} \tag{5.6}$$

With the class-based bigrams and trigrams, the probability of the current class c_i given the previous word(s) is backed off (again via uniform interpolation) to the probability given the previous class(es):

$$P_{\text{bi}}(c_i|w_{i-1}) = \frac{P(c_i|w_{i-1}) + P(c_i|c_{i-1})}{2} \tag{5.7}$$

$$P_{\text{tri}}(c_i|w_{i-2}^{i-1}) = \frac{P(c_i|w_{i-2}^{i-1}) + P(c_i|w_{i-2}, c_{i-1}) + P(c_i|c_{i-2}^{i-1})}{3} \tag{5.8}$$

With this model, varying λ varies the propensity to align.

5.4 Experiment 1: Repetition of Nouns and Adjectives

5.4.1 Introduction

5.4.1.1 Psycholinguistic Findings

Cleland and Pickering (2003) investigated the priming of noun-phrase structure in dialogue. Their first experiment was concerned with the repetition of nouns and adjectives between primes and targets. Using the confederate priming technique, participants described cards to each other. Each card showed one of 15 shapes (e.g., *circle*, *heart*, *square*) in one of 10 colours (e.g., *blue*, *green*, *orange*). The descriptions were categorised as *pre-nominal* (*the red square*), *relative clause* (*the square that 's red*) or *other*. From the counts, the *pre-nominal target ratio* was computed:

Pre-nominal target ratio The number of pre-nominal target responses divided by the sum of pre-nominal target responses and relative-clause target responses.

Pre-nominal target ratios were compared across eight different prime conditions:

- Pre-nominal, same noun, same adjective (between prime and target).
- Pre-nominal, different noun, same adjective.
- Pre-nominal, same noun, different adjective.
- Pre-nominal, different noun, different adjective.
- Relative clause, same noun, same adjective.
- Relative clause, different noun, same adjective.
- Relative clause, same noun, different adjective.
- Relative clause, different noun, different adjective.

Three-way analyses of variance (ANOVAs) with the independent variables *prime construction*, *noun* and *adjective* and participants and items as random effects were performed. There was a main effect of *prime construction*: Naïve participants were 19% more likely to repeat the construction of the confederate's prime utterance than to use the alternative. Furthermore, there was an interaction between *noun* and *prime*

construction: There was a 27% priming effect when prime and target used the same noun, as opposed to a 12% priming effect for differing nouns. Finally, there was a marginal interaction between *adjective* and *prime construction*, only significant across items: A 24% priming effect when prime and target used the same adjective, and a 13% priming effect for differing adjectives.

The prime conditions can be sorted by the pre-nominal target ratios which they elicit (according to Cleland and Pickering 2003, p. 220, Figure 3), starting with the highest pre-nominal target ratio:

1. Pre-nominal, same noun, same adjective.
2. Pre-nominal, same noun, different adjective.
3. Pre-nominal, different noun, same adjective.
4. Pre-nominal, different noun, different adjective.
5. Relative clause, different noun, different adjective.
6. Relative clause, different noun, same adjective.
7. Relative clause, same noun, different adjective.
8. Relative clause, same noun, same adjective.

This means that a *pre-nominal, same noun, same adjective* prime is most likely to elicit a pre-nominal target response, while a *relative clause, same noun, same adjective* prime is least likely to do so. If the list is read bottom-up, it shows the prime conditions most suited to elicit relative clause target responses.

5.4.1.2 Hypotheses

This suggests three hypotheses for our experiment, where we generate a prime phrase to initialise the cache language model and then a target phrase and observe the chosen edge's score:

1. The score for targets which keep the prime's syntactic construction should be higher than for those which do not.
2. The score for targets which keep the noun should be higher than for those which do not.

3. The score for targets which keep the adjective should be higher than for those which do not.

5.4.2 Methods

5.4.2.1 Materials

We picked six of the experimental items used by Cleland and Pickering, as shown in (5.9):

(5.9) (a) Red square/red diamond/green square/green diamond. Red square.

(b) Blue triangle/blue club/orange triangle/orange club. Blue triangle.

(c) Black circle/black spade/purple circle/purple spade. Black circle.

(d) Yellow sun/yellow heart/pink sun/pink heart. Yellow sun.

(e) Grey star/grey cross/brown star/brown cross. Grey star.

(f) Orange arrow/orange moon/pink arrow/pink moon. Orange arrow.

The four phrases separated by slashes serve as primes; the last phrase is the target. With respect to the target, the primes are presented in the order *same colour, same shape*/*same colour, different shape*/*different colour, same shape*/*different colour, different shape*. Care was taken not to choose the same colour or shape twice.

5.4.2.2 Procedure

The goal of the experiment was to replicate the priming effects found in Cleland and Pickering (2003) computationally by using cache language models. An OpenCCG grammar was written that was capable of parsing the materials in both pre-nominal and relative clause form. All possible realisations were then parsed and the resulting logical forms were saved in the OpenCCG XML format. These logical forms could then be used to re-generate the respective sentences as primes or targets.

For example, item (5.9)(a) yields parses of the following phrases:

(5.10) (a) the red square

(b) the square that 's red

(c) the red diamond

(d) the diamond that 's red

```
<?xml version="1.0" encoding="UTF-8"?>
<xml>
  <lf>
    <node id="s1:shape" pred="square" det="the" num="sg"
          tpc="+">
      <rel name="HasProp">
        <node id="r1:colour" pred="red" />
      </rel>
    </node>
  </lf>
  <target>the red square</target>
</xml>
```

Figure 5.1: OPENCCG logical form of the parse of phrase (5.10)(a).

(e) the green square

(f) the square that 's green

(g) the green diamond

(h) the diamond that 's green

The logical forms of the parses of phrases (5.10)(a) and (5.10)(b) are shown in Figures 5.1 and 5.2.

Each of the eight phrases (5.10)(a) to (5.10)(h) was then used as a prime for either the pre-nominal target (5.10)(a) or the relative clause target (5.10)(b). First the prime and then the target phrase were generated, which initialised the cache language model with the respective prime. The CLM's weight was modified from 0.0 to 1.0 in steps of 0.1 in order to observe the weight's effect; thus, each combination was scored 11 times.

Internally, OPENCCG sorted the edges by their geometric mean scores (see Section 4.3.2.1). As the grammar was small, OPENCCG pruning was switched off and all possible complete edges were generated, so the internal sorting had no influence on the scoring and it was not necessary to re-run the experiment with edge-sorting by probability. At the end, from all complete edges, the top-ranked GM score and the top-ranked probability were chosen as the results.

```
<?xml version="1.0" encoding="UTF-8"?>
<xml>
  <lf>
    <node id="s1:shape" pred="square" det="the" num="sg">
      <rel name="GenRel">
        <node id="b1:state" pred="be" tense="pres">
          <rel name="Arg">
            <node idref="s1:shape" />
          </rel>
          <rel name="Prop">
            <node id="r1:colour" pred="red">
              <rel name="Of">
                <node idref="s1:shape" />
              </rel>
            </node>
          </rel>
        </node>
      </rel>
    </node>
  </lf>
  <target>the square that 's red</target>
</xml>
```

Figure 5.2: OPENCCG logical form of the parse of phrase (5.10)(b).

Note that in order to compare our model to the psycholinguistic studies, we present our findings as parallel to theirs and compute scores for different primes given a fixed target. Equivalently, given a single prime and a set of paraphrases generated from a target logical form, the model predicts the probability of repeating a construction from an interlocutor's utterance, and that is what is used for the generation of dialogues in the following chapter.

5.4.2.3 Choice of Statistical Tests for Evaluation Purposes

Consequently, there were four sub-experiments: *Pre-nominal* or *relative clause* target combined with geometric mean scores or probabilities. Each of them was analysed with statistical tests to ascertain whether the prime condition had a significant effect on the mean scores.

The design of each sub-experiment comprises one dependent variable (*score*) and two independent variables or factors (*prime* and *CLM weight*). Factor *prime* has eight levels:

1. Pre-nominal, same noun, same adjective.
2. Pre-nominal, different noun, same adjective.
3. Pre-nominal, same noun, different adjective.
4. Pre-nominal, different noun, different adjective.
5. Relative clause, same noun, same adjective.
6. Relative clause, different noun, same adjective.
7. Relative clause, same noun, different adjective.
8. Relative clause, different noun, different adjective.

Factor *CLM weight* has 11 levels (the weight settings from 0.0 to 1.0 in steps of 0.1). In addition, there is a random factor *input sentence* which corresponds to the factor *subject* of other experiments; in our case, input sentences yielded scores by being rated with respect to varying conditions.

As scores of the same input sentences are observed for different primes and settings of the CLM weight, this is a two factor repeated measures (or within-subjects) design. Both *prime* and *CLM weight* are within-subject factors. A suitable statistical test for

this design is the repeated measures analysis of variance (ANOVA), the assumptions of which include:

- Normality: The data arise from populations with normal distribution.
- Homogeneity of variance: The variances of the assumed normal distributions are equal.
- Sphericity: The variances of the differences between all pairs of the repeated measurements are equal.

In the present case, though, it can neither be guaranteed that the geometric mean scores or probabilities are normally distributed, nor that the sphericity assumption is met.

A solution is to use the non-parametric equivalent of the one factor repeated measures ANOVA, the Friedman test for several related samples, which does not require knowledge of how the basic variables are distributed. Demšar (2006) suggests this methodology for comparisons of multiple classifiers.

The Friedman test assesses whether there is a significant overall difference between the mean ranks of the mean scores induced by the prime conditions. For each sub-experiment, in order to determine which differences between mean ranks were significant, seven post-hoc Wilcoxon signed-rank tests were performed, contrasting adjacent pairs. The Wilcoxon test is a non-parametric test for the comparison of two related samples; it is similar to the parametric t-test for paired samples. A Bonferroni correction was applied to the significance level α to account for the multiple comparisons, such that a p-value less than or equal to $0.05/7 = 0.0071$ was considered significant.

5.4.3 Results

The results of the four sub-experiments are presented in the following four subsections. In preparation for the Friedman test, the scores were averaged across the six experimental items (5.9)(a) to (5.9)(f) for each combination of prime and CLM weight. The corresponding graphs are shown below. For the graphs, we modify the prime conditions and keep the target's syntactic construction constant, which mirrors Cleland and Pickering's experimental design. The scores are plotted on a logarithmic scale because they cover a large range of values. Tables with the exact values of the means and standard deviations are available in Section B.1.1 in the appendix.

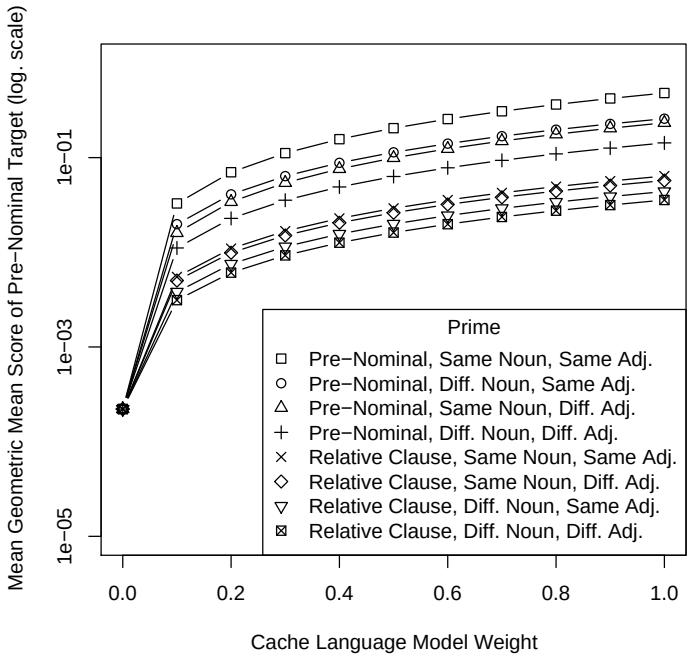

Figure 5.3: Mean geometric mean scores of pre-nominal targets following one of eight different primes.

The data of the extreme CLM weight settings 0.0 and 1.0 are excluded from the Friedman test: At CLM weight setting 0.0, the CLM is not taken into consideration and all scores are equal. On the other hand, at CLM weight setting 1.0, the base LM is not taken into consideration. If there happens to be a word in the target sentence that has neither a word match nor a semantic class match with the prime, the score calculation contains the factor 0 and the whole score ends up being 0. For example, given the prime *the red square* and the target *the square that 's red*, *that* and *'s* have no word or semantic class overlap with the prime, and they and consequently the target receive a score of 0 if the base LM does not contribute anything.

5.4.3.1 Geometric Mean Scores of Pre-Nominal Targets

Figure 5.3 shows the mean geometric mean scores of pre-nominal targets following one of eight different primes. As the CLM weight is increased, the mean GM scores in all prime conditions strictly increase.

A Friedman test was conducted to assess if there were differences among the mean

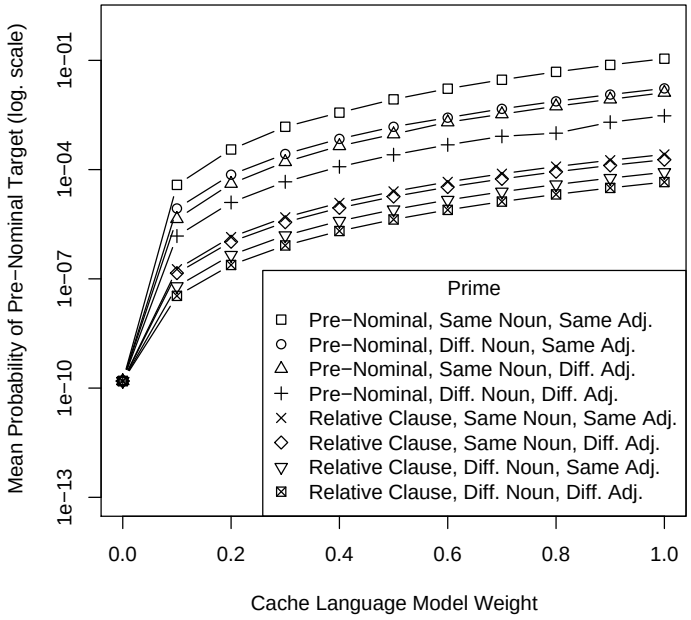

Figure 5.4: Mean probabilities of pre-nominal targets following one of eight different primes.

ranks of the mean GM scores induced by the prime conditions, $\chi_r^2(7, N = 9) = 63$, $p < 0.001$. This indicates that there were differences among the eight mean ranks. Seven orthogonal contrasts were performed using Wilcoxon tests with the Bonferroni correction (comparison-wise $\alpha = 0.0071$). All of these contrasts between prime conditions were found to be significant, $p = 0.0039$.

There is a clear effect of syntactic construction. The pre-nominal targets are assigned higher scores when following pre-nominal primes rather than relative clause primes. Within each construction condition, *same noun* and *same adjective* primes yield higher scores than primes where either the noun or the adjective was changed, which in turn yield higher scores than primes where both noun and adjective were changed.

5.4.3.2 Probabilities of Pre-Nominal Targets

In Figure 5.4, we see the mean probabilities of pre-nominal targets following one of eight different primes. The results are very similar to those in the condition with geo-

metric mean scores. With increasing CLM weight, the mean probabilities in all prime conditions strictly increase.

A Friedman test was conducted to assess if there were differences among the mean ranks of the mean probabilities induced by the prime conditions, $\chi_r^2(7, N = 9) = 63$, $p < 0.001$. This indicates that there were differences among the eight mean ranks. Seven orthogonal contrasts were performed using Wilcoxon tests with the Bonferroni correction (comparison-wise $\alpha = 0.0071$). All of these contrasts between prime conditions were found to be significant, $p = 0.0039$.

As with the GM scores, there is a clear effect of syntactic construction. The pre-nominal targets are assigned higher scores when following pre-nominal primes rather than relative clause primes. Within each construction condition, *same noun* and *same adjective* primes yield higher scores than primes where either the noun or the adjective was changed, which in turn yield higher scores than primes where both noun and adjective were changed.

5.4.3.3 Geometric Mean Scores of Relative Clause Targets

We now switch to relative clause targets. Figure 5.5 shows their mean geometric mean scores following one of eight different primes. As the CLM weight is increased, the mean GM scores in the relative clause prime conditions strictly increase. The mean GM scores in the pre-nominal prime conditions peak at a CLM weight of 0.6.

A Friedman test was conducted to assess if there were differences among the mean ranks of the mean probabilities induced by the prime conditions, $\chi_r^2(7, N = 9) = 63$, $p < 0.001$. This indicates that there were differences among the eight mean ranks. Seven orthogonal contrasts were performed using Wilcoxon tests with the Bonferroni correction (comparison-wise $\alpha = 0.0071$). All of these contrasts between prime conditions were found to be significant, $p = 0.0039$.

There is a clear effect of syntactic construction. The relative clause targets are assigned higher scores when following relative clause primes rather than pre-nominal primes. Within each construction condition, *same noun* and *same adjective* primes yield higher scores than primes where either the noun or the adjective was changed, which in turn yield higher scores than primes where both noun and adjective were changed.

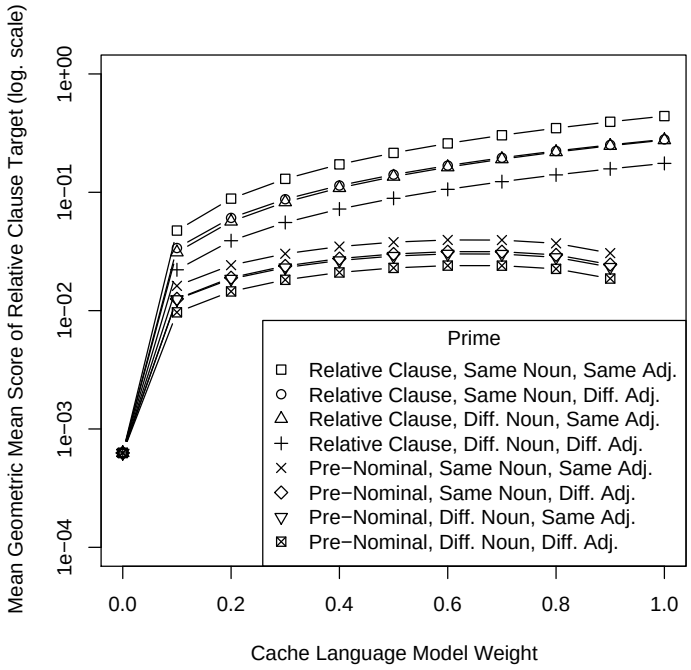

Figure 5.5: Mean geometric mean scores of relative clause targets following one of eight different primes.

5.4.3.4 Probabilities of Relative Clause Targets

Finally, Figure 5.6 shows the mean probabilities of relative clause targets following one of eight different primes. Similar to the GM score condition, with increasing CLM weight, the mean probabilities in the relative clause prime conditions strictly increase. The mean probabilities in the pre-nominal prime conditions peak at CLM weights of 0.6 in the *same adjective* conditions and 0.7 in the *different adjective* conditions.

A Friedman test was conducted to assess if there were differences among the mean ranks of the mean probabilities induced by the prime conditions, $\chi_r^2(7, N = 9) = 63$, $p < 0.001$. This indicates that there were differences among the eight mean ranks. Seven orthogonal contrasts were performed using Wilcoxon tests with the Bonferroni correction (comparison-wise $\alpha = 0.0071$). All of these contrasts between prime conditions were found to be significant, $p = 0.0039$.

Again, there is a clear effect of syntactic construction. The relative clause targets are assigned higher scores when following relative clause primes rather than pre-nominal primes. Within each construction condition, *same noun* and *same adjective*

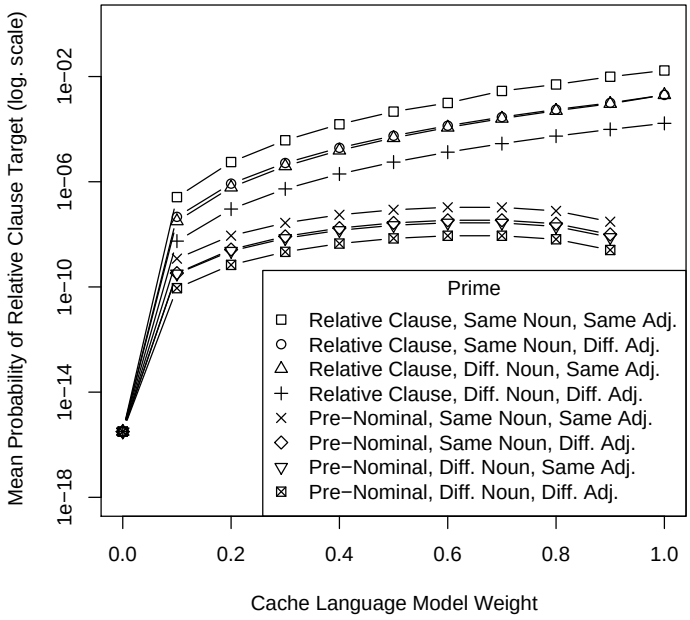

Figure 5.6: Mean probabilities of relative clause targets following one of eight different primes.

primes yield higher scores than primes where either the noun or the adjective was changed, which in turn yield higher scores than primes where both noun and adjective were changed.

5.4.4 Discussion

5.4.4.1 Hypotheses Revisited

The goal of this experiment was to explore in how far psycholinguistic findings on syntactic priming could be reproduced with the simple cache language modelling technique. We reiterate our hypotheses here:

1. The score for targets which keep the prime's syntactic construction should be higher than for those which do not.

2. The score for targets which keep the noun should be higher than for those which do not.

3. The score for targets which keep the adjective should be higher than for those which do not.

Hypothesis 1 was confirmed by the results. Scores for pre-nominal targets (e.g., *the red square*) were higher when the targets followed pre-nominal primes (e.g., *the red/green square/diamond*) than when they followed relative clause primes (e.g., *the square/diamond that 's red/green*). Similarly, scores for relative clause targets were higher when the targets followed relative clause primes than when they followed pre-nominal primes.

Hypotheses 2 and 3 were also confirmed. Keeping both noun and adjective (e.g., *red* and *square*) between prime and target yielded the highest scores. When only either the noun or the adjective differed, scores were lower than in the first case, but higher than when both differed.

5.4.4.2 Ordering by Target Ratios

According to Cleland and Pickering (2003), the prime conditions could be sorted by the pre-nominal target ratios which they elicited, starting with the highest pre-nominal target ratio:

1. Pre-nominal, same noun, same adjective.
2. Pre-nominal, same noun, different adjective.
3. Pre-nominal, different noun, same adjective.
4. Pre-nominal, different noun, different adjective.
5. Relative clause, different noun, different adjective.
6. Relative clause, different noun, same adjective.
7. Relative clause, same noun, different adjective.
8. Relative clause, same noun, same adjective.

With our setup, we cannot reproduce this order. We rather observe something like the following for the pre-nominal target condition:

1. Pre-nominal, same noun, same adjective.
2. Pre-nominal, different noun, same adjective.

3. Pre-nominal, same noun, different adjective.
4. Pre-nominal, different noun, different adjective.
5. Relative clause, same noun, same adjective.
6. Relative clause, same noun, different adjective.
7. Relative clause, different noun, same adjective.
8. Relative clause, different noun, different adjective.

At least the pre-nominal order is almost mirrored, while the relative clause condition is upside down: For humans, *relative clause, different noun, different adjective* primes yield more pre-nominal responses than *relative clause, same noun, same adjective* primes. With our model, it is the other way around.

Turning to the *relative clause target ratio*, Cleland and Pickering would predict this order:

1. Relative clause, same noun, same adjective.
2. Relative clause, same noun, different adjective.
3. Relative clause, different noun, same adjective.
4. Relative clause, different noun, different adjective.
5. Pre-nominal, different noun, different adjective.
6. Pre-nominal, different noun, same adjective.
7. Pre-nominal, same noun, different adjective.
8. Pre-nominal, same noun, same adjective.

Our model suggests:

1. Relative clause, same noun, same adjective.
2. Relative clause, same noun, different adjective.
3. Relative clause, different noun, same adjective.
4. Relative clause, different noun, different adjective.

5. Pre-nominal, same noun, same adjective.

6. Pre-nominal, same noun, different adjective.

7. Pre-nominal, different noun, same adjective.

8. Pre-nominal, different noun, different adjective.

As with pre-nominal targets, this only fits when the syntactic construction stays the same.

5.4.4.3 Overall Observations

While for most conditions, scores strictly increase when the CLM weight is increased, there is a peak at a weight setting of about 0.6 or 0.7 for relative clause targets with pre-nominal primes. An example for this would be the prime *the red square* with the target *the square that 's red*. This happens when the target contains words that have no word or semantic class overlap with the prime, related to the explanation for scores of 0 at a CLM weight setting of 1.0 (see Section 5.4.3). In our example, these words are *that* and *'s*. They only receive scores from the base language model, and when support from the base model is withdrawn by increasing the CLM weight, after a peak at a certain weight, the overall target score decreases again.

Consequently, to achieve maximum alignment across syntactic constructions, we suggest that the CLM weight should be set to 0.7, which is what we do for the experiments combining personality and lexical alignment in Chapter 6.

In this study, the CLM weight had no influence on the relative order of scores across prime conditions. However, in later experiments, we will encounter results where the relative order changes during the transition from low to high CLM weight.

Geometric mean scores and probabilities seem to behave similarly. Subsequently, we will therefore no longer consider probabilities but focus exclusively on GM scores instead.

The present experiment showed that it is possible to model syntactic priming effects with the cache language modelling approach. Even though some of the hypotheses do not match the psycholinguistic results exactly, overall, the model does what is expected from it: It causes a boost for similar-looking syntactic constructions by only relying on word sequences (with the fallback mechanism to semantic classes to avoid too many scores of 0 due to lack of overlapping words). We will now explore whether this also holds for psycholinguistic findings on semantic effects.

5.5 Experiment 2: Semantic Relatedness

5.5.1 Introduction

The design of this second experiment is very similar to that of the previous experiment. We will therefore not repeat all the details but only focus on the differences.

5.5.1.1 Psycholinguistic Findings

In a second study, Cleland and Pickering (2003) examined the influence of semantic relatedness on priming. This time, the participants described cards with everyday objects (e.g., *axe*, *cup*, *fence*) in different colours (e.g., *orange*, *pink*, *yellow*) to each other. Primes and targets either shared the head noun, or the head nouns were semantically related, or they were semantically unrelated. Again, the descriptions were categorised as *pre-nominal* (*the red sheep*), *relative clause* (*the sheep that 's red*) or *other*. Pre-nominal target ratios were compared across six different prime conditions:

- Pre-nominal, same noun (between prime and target).
- Pre-nominal, semantically related noun.
- Pre-nominal, semantically unrelated noun.
- Relative clause, same noun.
- Relative clause, semantically related noun.
- Relative clause, semantically unrelated noun.

Two-way analyses of variance (ANOVAs) with the independent variables *prime construction* and *semantic relatedness* and participants and items as random effects were performed. There was a main effect of *prime construction*: Naïve participants were 29% more likely to repeat the construction of the confederate's prime utterance than to use the alternative. Furthermore, there was an interaction between *prime construction* and *semantic relatedness*: There was a 47% priming effect when prime and target used the same noun, a 31% priming effect when prime and target used semantically related nouns, and an 8% priming effect when the nouns in prime and target were semantically unrelated.

The prime conditions can be sorted by the pre-nominal target ratios which they elicit (according to Cleland and Pickering 2003, p. 222, Figure 4), starting with the highest pre-nominal target ratio:

1. Pre-nominal, same noun.
2. Pre-nominal, semantically related noun.
3. Pre-nominal, semantically unrelated noun.
4. Relative clause, semantically unrelated noun.
5. Relative clause, semantically related noun.
6. Relative clause, same noun.

5.5.1.2 Hypotheses

This suggests three hypotheses for our experiment, where we once more generate a prime phrase to initialise the cache language model and then a target phrase and observe the chosen edge's score:

1. The score for targets which keep the prime's syntactic construction should be higher than for those which do not.
2. The score for targets which keep the noun should be higher than for those with a semantically related noun.
3. The score for targets with a semantically related noun should be higher than for those with a semantically unrelated noun.

5.5.2 Methods

5.5.2.1 Materials

This experiment re-uses six experimental items from Cleland and Pickering (2003)'s study on priming with semantically related nouns, as shown in (5.11):

(5.11) (a) Red sheep/red goat/red knife. Red sheep.

(b) Green axe/green saw/green cup. Green axe.

(c) Pink star/pink moon/pink fence. Pink star.

(d) Red tree/red bush/red boot. Red tree.

(e) Green arm/green leg/green bread. Green arm.

(f) Pink bed/pink cot/pink clock. Pink bed.

Per item, there are three primes and one target. The primes are given in the order *same noun*, *semantically related noun* and *semantically unrelated noun* with respect to the target, which follows at the end.

5.5.2.2 Procedure

Each item yields six phrases that were parsed with OPENCCG and stored as logical forms for later re-generation. For example, the phrases resulting from item (5.11)(a) are:

(5.12) (a) the red sheep

(b) the sheep that 's red

(c) the red goat

(d) the goat that 's red

(e) the red knife

(f) the knife that 's red

Each of the six phrases (5.12)(a) to (5.12)(f) was then used as a prime for either the pre-nominal target (5.12)(a) or the relative clause target (5.12)(b). First the prime and then the target phrase were generated, which initialised the cache language model with the respective prime. The CLM's weight was modified from 0.0 to 1.0 in steps of 0.1 in order to observe the weight's effect; thus, each combination was scored 11 times. At the end, from all complete edges, the top-ranked GM score was chosen as the result.

5.5.3 Results

There were two sub-experiments: *Pre-nominal* or *relative clause* targets in response to each of the six possible primes, only examining the geometric mean scores this time. Graphs with the results are presented in the following two subsections. Tables with the exact values of the means and standard deviations are available in Section B.1.2 in the appendix. For completeness, the appendix also contains tables with results for probabilities, but we will not discuss those any further.

As before, we applied a Friedman test to assess whether there was a significant overall difference between the mean ranks of the mean scores induced by the prime conditions. For each sub-experiment, in order to determine which differences between mean ranks were significant, five post-hoc Wilcoxon tests were performed, contrasting

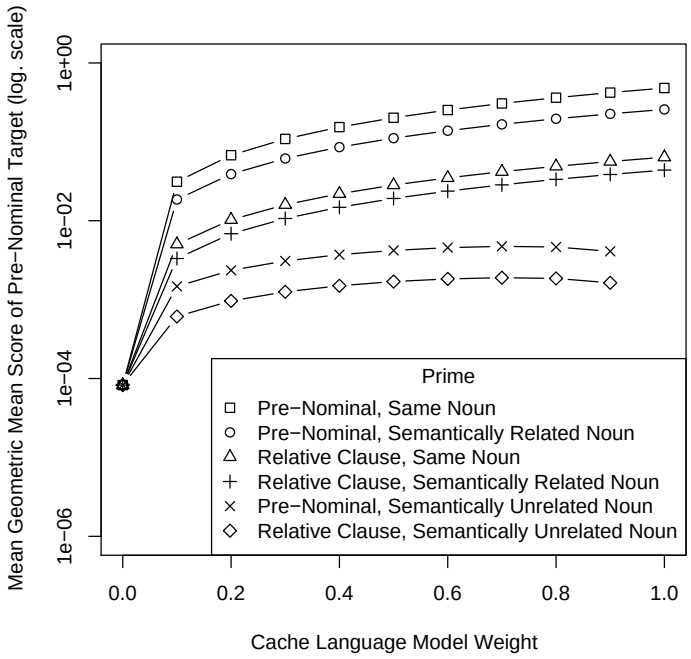

Figure 5.7: Mean geometric mean scores of pre-nominal targets following one of six different primes.

adjacent pairs. A Bonferroni correction was applied to the significance level α to account for the multiple comparisons, such that the p-value needed to be $0.05/5 = 0.01$ to be significant.

5.5.3.1 Pre-Nominal Targets

Figure 5.7 shows the mean geometric mean scores of *pre-nominal* targets following one of six different primes. With increasing CLM weight, the mean geometric mean scores in the *same noun* and *semantically related noun* prime conditions strictly increase; for primes with a *semantically unrelated noun*, they peak at CLM weight 0.7.

A Friedman test was conducted to assess if there were differences among the mean ranks of the mean GM scores induced by the prime conditions, $\chi^2_r(5, N = 9) = 45$, $p < 0.001$. This indicates that there were differences among the six mean ranks. Five orthogonal contrasts were performed using Wilcoxon tests with the Bonferroni correction (comparison-wise $\alpha = 0.01$). All of these contrasts between prime conditions were found to be significant, $p = 0.0039$.

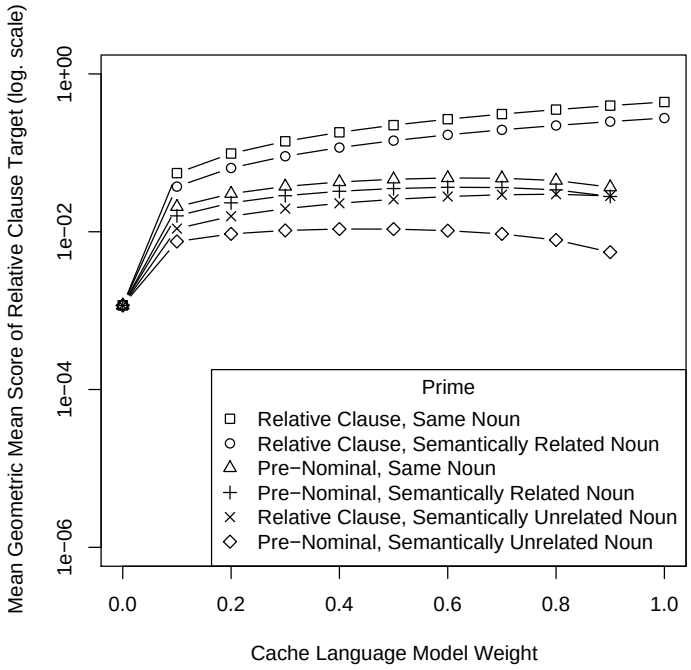

Figure 5.8: Mean geometric mean scores of relative clause targets following one of six different primes.

Pre-nominal primes with *same* or *semantically related nouns* yielded higher target scores than *relative clause* primes with *same* or *semantically related nouns*. The lowest scores were assigned to targets that were primed by *semantically unrelated nouns*, with the *pre-nominal* condition rated higher than the *relative clause* condition.

5.5.3.2 Relative Clause Targets

Figure 5.8 shows the mean geometric mean scores of *relative clause* targets following one of six different primes. For *relative clause* primes, with increasing CLM weight, the mean GM scores in the *same noun* and *semantically related noun* prime conditions strictly increase; in the *semantically unrelated noun* condition, they peak at CLM weight 0.8. For *pre-nominal* primes, mean GM scores peak at 0.6 in the *same noun* and *semantically related noun* prime conditions, and at 0.4 given *semantically unrelated noun* primes.

A Friedman test was conducted to assess if there were differences among the mean ranks of the mean GM scores induced by the prime conditions, $\chi_r^2(5, N = 9) = 44.49$,

$p < 0.001$. This indicates that there were differences among the six mean ranks. Five orthogonal contrasts were performed using Wilcoxon tests with the Bonferroni correction (comparison-wise $\alpha = 0.01$). All of these contrasts between prime conditions were found to be significant, with $p = 0.0078$ for the contrast between the *pre-nominal, semantically related noun* and the *relative clause, semantically unrelated noun* conditions, and $p = 0.0039$ for the other four.

Relative clause primes with *same* or *semantically related nouns* yielded higher target scores than *pre-nominal* primes with *same* or *semantically related nouns*. Similar to the results for *pre-nominal* targets, the lowest scores were assigned to targets that were primed by *semantically unrelated nouns*, this time with the *relative clause* condition rated higher than the *pre-nominal* condition.

5.5.4 Discussion

5.5.4.1 Hypotheses Revisited

This experiment examined whether it was possible to model semantic priming effects with the cache language modelling technique. Our hypotheses were:

1. The score for targets which keep the prime's syntactic construction should be higher than for those which do not.

2. The score for targets which keep the noun should be higher than for those with a semantically related noun.

3. The score for targets with a semantically related noun should be higher than for those with a semantically unrelated noun.

For Hypothesis 1, contrary to the results of the previous experiment, we no longer observe a clear separation that ranks target scores for *same construction* primes higher than target scores for *different construction* primes. As expected, the CLM's fallback to semantic classes has a boosting effect on the scores; both conditions with *semantically unrelated nouns* in the primes end up at the bottom of the ranking.

As far as Hypotheses 2 and 3 are concerned, they hold if we look at each of the prime syntactic constructions in isolation, i.e., within *pre-nominal* and *relative clause* primes we observe *same noun* > *semantically related noun* > *semantically unrelated noun*. Regarding the order based on all six primes, the only mismatch is that *same construction, semantically related noun* receives a higher score than *different construction, same noun*.

5.5.4.2 Ordering by Target Ratios

According to Cleland and Pickering (2003), the prime conditions could be sorted by the pre-nominal target ratios which they elicited, starting with the highest pre-nominal target ratio:

1. Pre-nominal, same noun.
2. Pre-nominal, semantically related noun.
3. Pre-nominal, semantically unrelated noun.
4. Relative clause, semantically unrelated noun.
5. Relative clause, semantically related noun.
6. Relative clause, same noun.

With the cache language modelling technique, we observe:

1. Pre-nominal, same noun.
2. Pre-nominal, semantically related noun.
3. Relative clause, same noun.
4. Relative clause, semantically related noun.
5. Pre-nominal, semantically unrelated noun.
6. Relative clause, semantically unrelated noun.

The order only matches for the first two items. The *relative clause, same noun* condition was moved up from the sixth to the third position and the *relative clause, semantically related noun* condition was moved up from the fifth to the fourth position.

Sorted by relative clause target ratio, the psycholinguistic results would predict this order:

1. Relative clause, same noun.
2. Relative clause, semantically related noun.
3. Relative clause, semantically unrelated noun.
4. Pre-nominal, semantically unrelated noun.

5. Pre-nominal, semantically related noun.

6. Pre-nominal, same noun.

Our model yields:

1. Relative clause, same noun.

2. Relative clause, semantically related noun.

3. Pre-nominal, same noun.

4. Pre-nominal, semantically related noun.

5. Relative clause, semantically unrelated noun.

6. Pre-nominal, semantically unrelated noun.

Once more, there is an overlap only for the first two items. The *pre-nominal, same noun* condition was moved up from the sixth to the third position and the *pre-nominal, semantically related noun* condition was moved up from the fifth to the fourth position.

5.5.4.3 Overall Observations

In this experiment, the curves that do not strictly increase peak twice at 0.7 for *pre-nominal* targets and at 0.4, 0.8 and twice at 0.6 for *relative clause* targets. Apart from the outlier at 0.4, this lends further support for a CLM weight setting of 0.7 to achieve maximum alignment, as was already proposed in Section 5.4.4.3.

After having examined the behaviour of cache language models for the pre-nominal/relative clause alternation, either without or with semantic effects, we now describe a third and final experiment that deals with a different syntactic structure, the prepositional object/double object alternation, and also with the effect of keeping the verb between prime and target.

5.6 Experiment 3: Prepositional Object/Double Object Priming

5.6.1 Introduction

5.6.1.1 Psycholinguistic Findings

Pickering and Branigan (1998) investigated the priming of prepositional object (PO)

		Target Completion	
Verb Type	Prime Completion	PO	DO
Same	PO	0.47	0.22
	DO	0.29	0.38
Different	PO	0.40	0.25
	DO	0.35	0.29

Table 5.2: Proportions of prepositional object (PO) and double object (DO) target responses following PO or DO prime completions. Results of Experiment 1 from Pickering and Branigan (1998, p. 639, Table 1).

and double object (DO) constructions. We focus on the results of their first experiment. Participants were given a booklet in which they were asked to complete sentence fragments. There were prime fragments which specifically induced PO or DO completions, followed by target fragments which could be completed as either construction. The verb between prime and target fragment either remained the same or was changed. Additionally, there were filler fragments unrelated to the topic of the investigation, none of which contained a verb that could be completed with a PO or DO construction:

- Noun phrases of varying types, including some containing verbs in embedded clauses.
- Noun phrases followed by a verb.
- Noun phrases followed by a verb and a noun phrase.

Pickering and Branigan computed the proportions of PO and DO target responses following PO prime completions, and those following DO prime completions. The results are reproduced in Table 5.2.

Three-way analyses of variance (ANOVAs) with the independent variables *prime completion*, *target completion* and *verb type* were performed. There was an interaction of *prime completion* and *target completion*: Participants produced significantly more target completions that exhibited the same syntactic construction as the prime comple-

tions than target completions that switched the syntactic construction. Overall, 11.7% more *PO* targets followed *PO* primes than *DO* primes, and 9.8% more *DO* targets followed *DO* primes than *PO* primes.

The analysis of variance (ANOVA) also revealed a three-way interaction of *prime completion*, *target completion* and *verb type*: When the verb remained the same between prime and target, participants produced 17.2% more same-type than alternative-type target completions, while when the verb varied, they only produced 4.4% more same-type than alternative-type target completions. Thirdly, there was a weak tendency for participants to produce more *PO* than *DO* completions.

The prime conditions can be sorted by the proportions of *PO* target responses which they elicit (according to Table 5.2), starting with the highest proportion:

1. PO, same verb.
2. PO, different verb.
3. DO, different verb.
4. DO, same verb.

Sorting by the proportions of *DO* target responses yields this list in reverse order:

1. DO, same verb.
2. DO, different verb.
3. PO, different verb.
4. PO, same verb.

5.6.1.2 Hypotheses

The psycholinguistic findings suggest two hypotheses for our experiment, where we, as before, generate a prime phrase to initialise the cache language model and then generate a target phrase and observe the chosen edge's score:

1. The score for targets which keep the prime's syntactic construction should be higher than for those which do not.
2. The score for targets which keep the verb should be higher than for those with a different verb.

5.6.2 Methods

5.6.2.1 Materials

We randomly chose verbs and nouns from the materials in the appendix of Pickering and Branigan (1998) to create a list of six experimental items for use in our experiment, taking care to avoid repetition as much as possible:

(5.13) (a) The secretary handed/sent the fax/the businessman. The grandmother handed the present/the girl.

(b) The captain gave/lent the lifejacket/the sailor. The student gave the money/the friend.

(c) The millionaire loaned/gave the painting/the artist. The swimmer loaned the towel/the diver.

(d) The researcher sent/posted the results/the surgeon. The photographer sent the prints/the editor.

(e) The child showed/gave the book/the friend. The architect showed the plans/the engineer.

(f) The hostess offered/handed the dessert/the guests. The barman offered the cocktail/the customer.

Primes were constructed from the first part of each item: The verb before the slash was used in the *same verb* prime conditions; the verb after the slash was used in the *different verb* prime conditions. Targets were constructed from the second part of each item. Both primes and targets were used in either their *PO* or their *DO* variants.

5.6.2.2 Procedure

Each item yields six phrases that were parsed with OPENCCG and stored as logical forms for later re-generation. For example, the phrases resulting from item (5.13)(a) are:

(5.14) (a) the secretary handed the fax to the businessman

(b) the secretary handed the businessman the fax

(c) the secretary sent the fax to the businessman

(d) the secretary sent the businessman the fax

(e) the grandmother handed the present to the girl

(f) the grandmother handed the girl the present

Each of the four phrases (5.14)(a) to (5.14)(d) was then used as a prime for either the *PO* target (5.14)(e) or the *DO* target (5.14)(f). First the prime and then the target phrase were generated, which initialised the cache language model with the respective prime. The CLM's weight was modified from 0.0 to 1.0 in steps of 0.1 in order to observe the weight's effect; thus, each combination was scored 11 times. At the end, from all complete edges, the top-ranked GM score was chosen as the result.

5.6.3 Results

There were two sub-experiments: *PO* or *DO* targets in response to each of the four possible primes, once more only examining the geometric mean scores. Graphs with the results are presented in the following two subsections. Tables with the exact values of the means and standard deviations are available in Section B.1.3 in the appendix. For completeness, the appendix also contains tables with results for probabilities, but we will not discuss those any further.

As before, we applied a Friedman test to assess whether there was a significant overall difference between the mean ranks of the mean scores induced by the prime conditions. For each sub-experiment, in order to determine which differences between mean ranks were significant, three post-hoc Wilcoxon tests were performed, contrasting adjacent pairs. A Bonferroni correction was applied to the significance level α to account for the multiple comparisons, such that the p-value needed to be $0.05/3 = 0.0167$ to be significant.

5.6.3.1 Prepositional Object Targets

Figure 5.9 shows the mean geometric mean scores of *prepositional object* targets following one of four different primes. With increasing CLM weight, the mean geometric mean scores in the *PO* prime condition strictly increase up to a weight of 0.9; for weight 1.0, the scores drop sharply. Scores for targets of primes with a *DO* construction peak at CLM weight 0.8.

A Friedman test was conducted to assess if there were differences among the mean ranks of the mean GM scores induced by the prime conditions, $\chi_r^2(3, N = 9) = 25.9333$, $p < 0.001$. This indicates that there were differences among the four mean

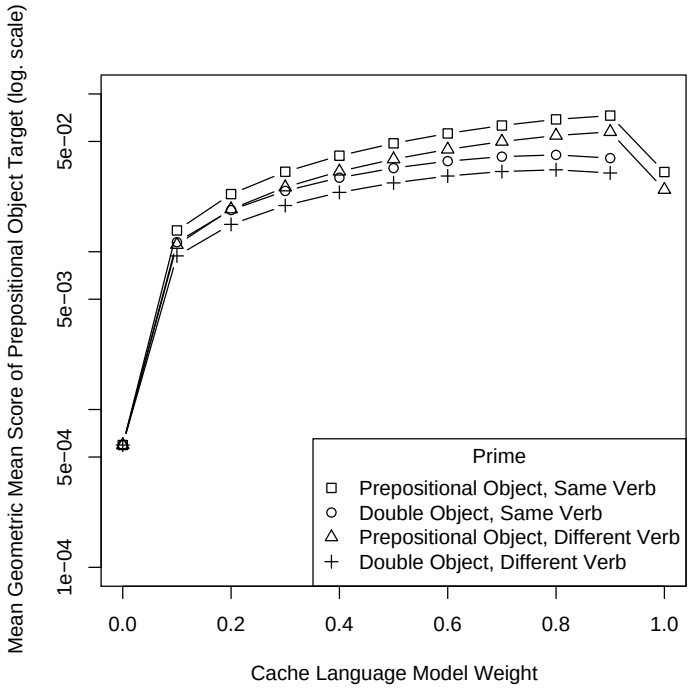

Figure 5.9: Mean geometric mean scores of prepositional object targets following one of four different primes.

ranks. Three orthogonal contrasts were performed using Wilcoxon tests with the Bonferroni correction (comparison-wise $\alpha = 0.0167$). All of these contrasts between prime conditions were found to be significant, with $p = 0.0117$ for the contrast between *PO, different verb* and *DO, same verb*, and $p = 0.0039$ for the other two contrasts.

The lines for the *PO, different verb* and *DO, same verb* prime conditions cross at a point between the CLM weight settings of 0.1 and 0.2. After that, the order stabilises and we observe that targets following *PO* primes received higher scores than targets following *DO* primes, and within the *PO*/*DO* conditions, *same verb* primes yielded higher scores than *different verb* primes.

5.6.3.2 Double Object Targets

Figure 5.10 shows the mean geometric mean scores of *double object* targets following one of four different primes. In all prime conditions, when the CLM weight is increased, the target scores strictly increase up to a CLM weight of 0.9, after which they drop sharply, just like the *PO* prime, *PO* target scores reported above.

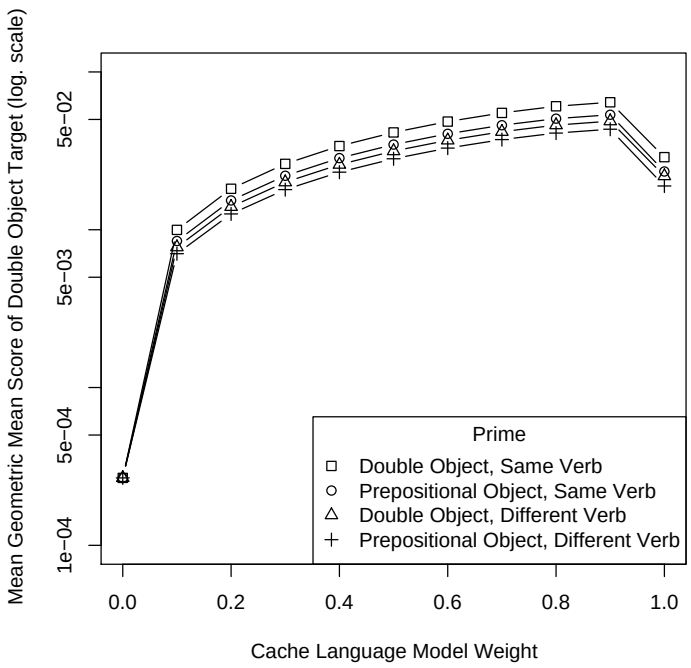

Figure 5.10: Mean geometric mean scores of double object targets following one of four different primes.

A Friedman test was conducted to assess if there were differences among the mean ranks of the mean GM scores induced by the prime conditions, $\chi_r^2(3, N = 9) = 27$, $p < 0.001$. This indicates that there were differences among the four mean ranks. Three orthogonal contrasts were performed using Wilcoxon tests with the Bonferroni correction (comparison-wise $\alpha = 0.0167$). All of these contrasts between prime conditions were found to be significant, $p = 0.0039$.

Targets following *same verb* primes received higher scores than targets following *different verb* primes, and within the *same verb*/*different verb* conditions, *DO* primes yielded higher scores than *PO* primes.

5.6.4 Discussion

5.6.4.1 Hypotheses Revisited

This experiment examined whether it was possible to use the cache language modelling technique to model the priming of PO/DO structures and the effect of keeping or

switching the verb between prime and target. Our hypotheses were:

1. The score for targets which keep the prime's syntactic construction should be higher than for those which do not.

2. The score for targets which keep the verb should be higher than for those with a different verb.

Both hypotheses were confirmed. With *PO* targets, *PO* primes induced higher scores than *DO* primes. Within the *prime completion* condition, *same verb* primes yielded higher scores than *different verb* primes. On the other hand, with *DO* targets, *same verb* primes induced higher scores than *different verb* primes. Within the *verb type* condition, *DO* primes yielded higher scores than *PO* primes.

5.6.4.2 Ordering by Target Ratios

According to Pickering and Branigan (1998), the prime conditions could be sorted by the proportions of *PO* target responses which they elicited, starting with the highest proportion:

1. PO, same verb.

2. PO, different verb.

3. DO, different verb.

4. DO, same verb.

With the cache language modelling technique, we observe:

1. PO, same verb.

2. PO, different verb.

3. DO, same verb.

4. DO, different verb.

The orders match for the first two items. The *DO, same verb* condition was moved up from the fourth to the third position.

Sorted by the proportions of *DO* target responses, the psycholinguistic findings indicate this order:

1. DO, same verb.
2. DO, different verb.
3. PO, different verb.
4. PO, same verb.

Our model produces:

1. DO, same verb.
2. PO, same verb.
3. DO, different verb.
4. PO, different verb.

Only the top item is sorted in accordance with human behaviour. The *PO, same verb* condition was moved up from the fourth to the second position.

In both cases, the CLM boosts the *different prime completion, same verb* condition due to the overlapping verb.

5.6.4.3 Overall Observations

Figure 5.9 shows that the order of target scores stabilises at higher CLM weight settings, and the lines for the scores of targets following *DO* primes peak at a CLM weight of 0.8. Looking at all three experiments, the peaks for the respective conditions lie between 0.6 and 0.8 (apart from one outlier at 0.4 in the second experiment), so 0.7 seems to be a good overall setting to achieve maximum alignment, as already suggested in Sections 5.4.4.3 and 5.5.4.3.

5.7 Discussion

In this chapter, we suggested cache language models as a lexically-based computational model of a type of matching. This approach approximates the main elements of human performance in psycholinguistic experiments on priming and alignment.

For each of the three experiments, the hypotheses we formulated about the relative order of target scores were mostly confirmed. The model cannot reproduce the

exact order that would be expected from a human participant, but the top-ranked target always matches, and in the second experiment, there is an overlap for the top two targets.

The fallback to semantic classes when there is no word match has a boosting effect that leaves utterances with neither word nor semantic class match at the bottom of the ranking, which does not happen with human judgements.

An advantage of the model is its simplicity and ease of integration with a natural language generation system, in this case the OPENCCG realiser. As we want to maximise alignment behaviour in our subsequent experiment, it is perfectly fine if only the top-ranked utterance corresponds to human performance, as this is the one we will choose to output in the dialogues we generate. To this end, the experiments in this chapter helped to determine a weighting for the CLM which maximises the matching effect.

The goal of this chapter was to explore the capabilities of simple CLMs. In future work, the current word-based approach could be extended with further features, e.g., richer part-of-speech tags, or *supertags* (Bangalore and Joshi 1999). With CCG (and related lexicalised theories of syntax), supertags encode the syntactic category of a lexical item. For example, a verb's annotation would then already give an indication of what kind of arguments it expects. This information provides an extended domain of locality, and the model would be less dependent on exact word sequences. Techniques from supertagging have recently been integrated into OPENCCG as *hypertagging* and have been shown to improve realisation speed and quality (Espinosa et al. 2008).

Taking a broader view, what use is this type of matching? Brennan (1996) studied *lexical entrainment*. During a conversation, participants form 'conceptual pacts'. They reduce the high variability in lexical choices by agreeing on referring expressions. Brennan also found lexical convergence of users with computers. People were at least as likely to adopt the terms of their computer partners as those of their human partners.

Pearson et al. (2006) focused on human–computer interaction and found that users adapt their language according to their expectations about a system's capabilities. In a picture-naming and -matching game, participants interacted with a *basic* and an *advanced* version of a computer program that differed in the messages displayed on the start-up screen, but behaved identically otherwise. Alignment was significantly greater than chance in both conditions and significantly greater in the *basic* condition than in the *advanced* condition.

So, evidence for alignment has been found between humans as well as between

humans and computers. But how do people perceive computer–computer interactions which exhibit alignment behaviour? This is the topic of Chapter 6, in which we consider the effect of matching, generated with the CLM approach, in comparison with a lack of matching, where the interlocutors differ in linguistic style according to their personality as they did in Chapter 4.

Chapter 6

A Unified Model of Personality and Alignment

6.1 Summary

After optimal parameters to project alignment were determined through the experiments in the previous chapter, this chapter describes and evaluates the Affective Language Production Model version 4 (ALPM-4), which augments the ALPM-3 of Chapter 4 with alignment by way of cache language models. With this approach, it is possible to examine dynamic, short-term alignment effects in direct contrast with stable, long-term personality effects. This combines the two contrasting sources of individual differences in language use studied throughout this thesis. Generated dialogues were evaluated in a further condition of the web-based experiment that was conducted in Chapter 4, allowing for direct comparison of the perception of dialogues with and without alignment. We found that the introduction of alignment significantly reduced the agreement between judges and the model for the agreeableness dimension. In addition to that, the computer characters were perceived to be getting on less well with each other and their dialogues were rated as less smooth.

The approach discussed in this chapter uses the implementations of the utterance ranking and cache language model components already described in Chapters 4 and 5. Further evaluation was carried out by Carsten Brockmann.

6.2 Introduction

Niederhoffer and Pennebaker (2002) expected that an interaction among relative stran-

gers would exhibit comparable word use if it went well, and that conflict within the interaction would manifest itself in unmatched patterns of language between the dialogue participants. However, they found no relationship between perceived interaction quality and linguistic style matching, neither for self-report nor for judges' ratings. To explain these results, they suggested a *coordination-engagement* hypothesis:

> ..., the more that two people in a conversation are actively engaged with one another—in a positive or even negative way—the more verbal and nonverbal coordination we expect.

This means that even if the dialogue participants disagree, there should be alignment as long as they are engaged in the conversation.

In this chapter, we propose the Affective Language Production Model version 4 (ALPM-4), which unifies the treatment of personality and alignment processes in dialogue. We assess and examine the effect of alignment behaviour on the perception of personality and interaction quality in comparison with the results obtained with ALPM-3.

6.2.1 Hypotheses

We hypothesise that introducing alignment via ALPM-4 makes it more difficult for the participants to properly recognise the characters' personalities, due to the conflict between the dynamic, short-term alignment effects and the stable, long-term personality effects on language behaviour; in a way, we expect alignment to overwrite the personality effects. This should manifest itself in reduced percentages of agreement between the model's settings and the judges' ratings.

We also hypothesise that alignment between the dialogue participants has an effect on the perception of how well the characters got on with each other and how smoothly the conversation went. The hypothesis is two-tailed; the effect might be positive or negative.

6.3 Methods

6.3.1 Natural Language Generation Guided by Cache Language Models

ALPM-4 uses the same framework as ALPM-3 (see Section 4.3). There is an additional component that takes care of modelling alignment via cache language models (CLMs),

		Personality Parameter Setting					Propensity to
Condition	Character	E	N	A	C	O	Align
A) High E	A	75	50	25	25	50	0
vs. Low E	B	25	50	75	75	50	0.7
B) Low E	A	25	50	25	25	50	0
vs. High E	B	75	50	75	75	50	0.7
C) High N	A	50	75	25	25	50	0
vs. Low N	B	50	25	75	75	50	0.7
D) Low N	A	50	25	25	25	50	0
vs. High N	B	50	75	75	75	50	0.7

Table 6.1: Personality and alignment parameter settings for the four experimental conditions.

with the technique introduced in the previous chapter (see Section 5.3). A CLM is computed based on the utterance that was generated immediately before. The CLM is then combined with the personality LM that was already used in ALPM-3. A character's propensity to align corresponds to the weight given to the CLM during this combination, and can thus be set to a value between 0 and 1.

6.3.2 Materials

6.3.2.1 Generation

To be able to compare human judges' perceptions of dialogues with and without alignment, we took the materials generated for the ALPM-3 experiment as a basis (see Section 4.3.6.3). The dialogues had been generated in four different conditions, as shown in Table 6.1. Each condition sets the two computer characters to opposing extremes on either the E or the N dimension, while keeping the respective other dimension at a middle level. Furthermore, character A is always Low A and C, and character B is always High A and C. All characters are set to Mid O.

The difference to the previous experiment is that alignment is switched on for the

High A/High C characters, in parallel to the ALPM-2 assumption that characters low in psychoticism would have a propensity to align (see Section 3.3.3), and linked to the ALPM-3 topic choice strategy (see Section 4.3.5). There is empirical evidence which suggests that a moderate level of neuroticism facilitates priming and a high level of N inhibits it (Gill et al. 2004), but this has yet to be replicated on a larger sample. The weight for the cache language model is set to 0.7, according to the findings in Chapter 5.

Utterances for the non-aligning speaker were kept as before. The generation of utterances for the aligning speaker was seeded with the respective previous utterance functioning as the dialogue history. From the list of generated utterances, the top-ranked utterance was chosen.

6.3.2.2 Web Experiment

The experiment was conducted on the World Wide Web, together with the experiment evaluating ALPM-3. See Section 4.3.6.4 for a description of the introductory page and Table 4.6 for the exemplar personality dimension reminders displayed at the bottom of each dialogue.

The participants were presented with one dialogue per condition, randomly chosen from the pool of available dialogues, in random order (mixed with the stimuli of the other experiment). One dialogue was shown at a time, and the participants were asked to judge which of the two characters scored higher on the E, N, A and C dimensions. The higher-scoring character had to be chosen; *equal* was not an option. The judges also assessed, on a seven-point Likert scale, how well the characters got on with each other and how smoothly the conversation went, with one point representing *very badly* or *not at all smoothly* and seven points representing *very well* or *very smoothly*, respectively. The participants were asked to rate each dialogue independently from the others.

6.3.2.3 Example Dialogues

Here we show the same dialogues as in Section 4.3.6.5, this time with alignment switched on for the High A/High C characters (Character B in both cases). Character A's utterances are the same as before. Character B's previous, non-aligning utterances are shown in brackets to demonstrate what effect the alignment has.

The aligned utterances closely match the surface form of their primes. There is a

recognisable difference to the utterances chosen in the non-aligning version.

6.3.2.3.1 High E versus Low E

Parameter Settings

Character A's personality High E, Mid N, Low A, Low C, Mid O.

Character A's propensity to align 0.

Character A's agenda PLOT/−, CHARACTERS/−, MUSIC/−, FILM/−.

Character A's further opinions SPECIAL EFFECTS/−, ACTION SCENES/+, DIALOGUE/−.

Character B's personality Low E, Mid N, High A, High C, Mid O.

Character B's propensity to align 0.7.

Character B's agenda ACTION SCENES/+, SPECIAL EFFECTS/+, PLOT/−, DIALOGUE/−.

Character B's further opinions CHARACTERS/−, FILM/−, MUSIC/−.

Movie Mystic River.

Base language model SWITCHBOARD LM, weight 0.5; CrAg Corpus LM, weight 0.5.

Combined base and personality language model Base LM, weight 0.3; personality LM, weight 0.7.

Dialogue

A (PLOT) "I mean I didn't think that the story was great."

B (PLOT) "I mean the story wasn't great."
[*Previously:* "Yes I mean I thought that it wasn't nicely tied up."]

A (CHARACTERS) "And I mean I thought there was no characterization at all."

B (CHARACTERS) "I mean I thought there was no characterization at all."
[*Previously:* "Yeah there was no characterization at all I mean."]

A (MUSIC) "I mean I didn't think that the music was great."

B (MUSIC) "I mean the music wasn't great."
[*Previously:* "Yeah I kind of thought that the music was also very bad you know."]

A (FILM) "I mean I didn't think the film was good."

B (FILM) "I mean the film wasn't good."
[*Previously:* "I mean I kind of thought that it was a bit dull."]

6.3.2.3.2 High N versus Low N

Parameter Settings

Character A's personality Mid E, High N, Low A, Low C, Mid O.

Character A's propensity to align 0.

Character A's agenda DIALOGUE/−, FILM/−, PLOT/−, ACTION SCENES/+.

Character A's further opinions MUSIC/−, CHARACTERS/+, SPECIAL EFFECTS/+.

Character B's personality Mid E, Low N, High A, High C, Mid O.

Character B's propensity to align 0.7.

Character B's agenda ACTION SCENES/+, PLOT/+, SPECIAL EFFECTS/−, CHARACTERS/+.

Character B's further opinions MUSIC/+, DIALOGUE/+, FILM/+.

Movie Intolerable Cruelty.

Base language model SWITCHBOARD LM, weight 0.5; CrAg Corpus LM, weight 0.5.

Combined base and personality language model Base LM, weight 0.3; personality LM, weight 0.7.

Dialogue

A (DIALOGUE) "So I didn't think that the script was very good you know."

B (DIALOGUE) "The script was very good you know."
[*Previously:* "Well actually to be honest I didn't think that the script was very bad."]

A (FILM) "I mean I didn't think the film was very good."

B (FILM) "I mean the film was very good."
[*Previously:* "Well actually I mean I didn't think that the movie was very bad."]

A (PLOT) "But you know I didn't think that the story was very good."

B (PLOT) "The story was very good."
[*Previously:* "I mean I thought the story was really good."]

A (ACTION SCENES) "I mean I didn't think that the action scenes were very bad really."

B (ACTION SCENES) "I mean the action scenes were very good."
[*Previously:* "Yeah I kind of thought that the action scenes were very good also you know."]

6.4 Results

The same participant statistics as described in Section 4.4 apply. We repeat them here for convenience: Eighty participants completed the web experiment. The data were filtered to exclude replies that were submitted after less than five minutes (five cases) or more than 45 minutes (one case), leaving 74 judgements for further analysis. Of these, 39 were native and 35 were non-native speakers of English.

The data were evaluated by comparing the settings used to generate the experimental materials with the actual human judgements. The first hypothesis was that for conditions A and B people would be able to detect differences in extraversion, while for conditions C and D they would see differences in neuroticism. Differences in agreeableness and conscientiousness were expected across all four conditions. Table 6.2 shows the results of this comparison. Figures 6.1 to 6.3 show the percentage of participants' choices in agreement with expectations for the native, non-native and all speakers conditions, respectively.

	Native Speakers				Non-Native Speakers				All			
Choice	E	N	A	C	E	N	A	C	E	N	A	C
Expected	40	44	70	77	41	37	62	69	81	81	132	146
Not expected	38	34	86	79	29	33	78	71	67	67	164	150
Number of trials	78	78	156	156	70	70	140	140	148	148	296	296
Expected (%)	51.3	56.4	44.9	49.4	58.6	52.9	44.3	49.3	54.7	54.7	44.6	49.3
Significance of binomial test												
Expected without alignment (%)	55.1	59.0	68.6	52.6	47.1	58.6	70.0	39.3	51.4	58.8	69.3	46.3
Difference with alignment (%)	−3.8	−2.6	−23.7	−3.2	+11.5	−5.7	−25.7	+10.0	+3.3	−4.1	−24.7	+3.0
Significance of chi-square test			$***$				$***$				$***$	

Table 6.2: Numbers of expected versus not expected choices for the web experiment, with results of binomial tests to determine whether the distributions differ from chance. The bottom of the table shows the percentage differences to the results without alignment (cf. Table 4.7) and the results of chi-square tests to determine whether the differences are statistically significant. Significance levels of statistics: $*$: $p \leq 0.05$, $**$: $p \leq 0.01$, $***$: $p \leq 0.001$.

Binomial tests were performed to assess whether the distribution of values differed significantly from chance. Native English speakers agreed with ALPM-4 more than 50% of the time for the E and N dimensions; the overall order was N > E > C > A. Non-native speakers also agreed with the model more than 50% of the time for the E and N dimensions, but the percentage for E was higher and the percentage for N was lower than for the native speakers; the order was E > N > C > A. Consequently, all participants taken as a group agreed with ALPM-4 more than 50% of the time for the E and N dimensions; the overall percentage turned out to be exactly equal, leading to the order E = N > C > A. None of the percentages differed significantly from chance, though.

With the introduction of alignment, the personality dimensions were expected to be more difficult to recognize than before. Table 6.2's bottom rows and Figures 6.4 to 6.6 show both conditions in comparison. With alignment switched on, native speakers agreed with the model less for all four personality dimensions; slightly less for E (-3.8%), N (-2.6%) and C (-3.2%), and much less for A (-23.7%). Non-native speakers agreed more for E ($+11.5\%$) and C ($+10.0\%$), less for N (-5.7%) and much less for A (-25.7%). Overall, the participants agreed slightly more for E ($+3.3\%$) and C ($+3.0\%$), slightly less for N (-4.1%) and much less for A (-25.7%).

Pearson's chi-square tests were computed to assess whether alignment had an effect on the distribution of expected and not expected choices. A significant difference was found only for the A personality dimension, across all three groups of participants: Native ($\chi^2 = 16.922$, $df = 1$, $N = 312$, $p \leq 0.001$), non-native ($\chi^2 = 17.865$, $df = 1$, $N = 280$, $p \leq 0.001$) and all speakers ($\chi^2 = 35.712$, $df = 1$, $N = 592$, $p \leq 0.001$). The non-aligning condition is more likely than expected under the null hypothesis to elicit higher agreeableness ratings from the judges than the aligning condition.

Figures 6.7 to 6.9 show the distribution of ratings of how well the characters got on with each other and how smoothly the conversation went, for native, non-native and all speakers, respectively. The average native speakers' judgements for getting on (median $m = 4$, mean $\bar{x} = 3.590$, standard deviation $s = 1.650$) and smoothness ($m = 3$, $\bar{x} = 3.340$, $s = 1.660$) were slightly lower than the non-native judgements for getting on ($m = 4$, $\bar{x} = 3.736$, $s = 1.665$) and smoothness ($m = 4$, $\bar{x} = 3.971$, $s = 1.666$), respectively. The difference between means for getting on (0.146) was lower than for smoothness (0.631). The overall values for getting on ($m = 4$, $\bar{x} = 3.659$, $s = 1.656$) and smoothness ($m = 3$, $\bar{x} = 3.639$, $s = 1.690$) lay in between.

Table 6.3 and Figures 6.10 to 6.12 show a direct comparison of the getting on and

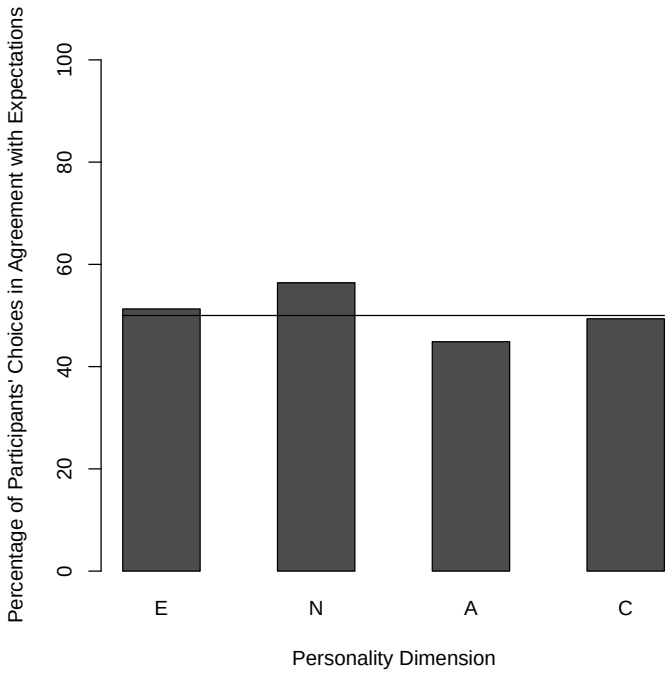

Figure 6.1: Percentage of native English speakers' choices in agreement with expectations, by personality dimension, with alignment.

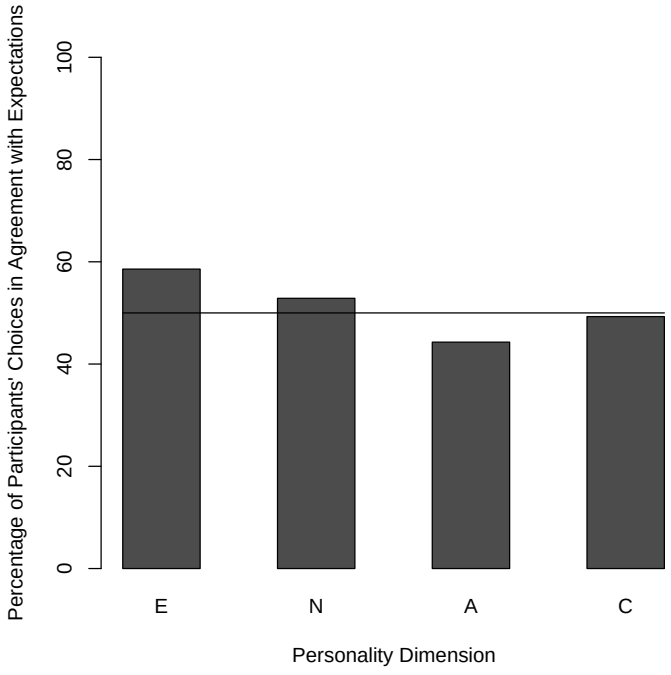

Figure 6.2: Percentage of non-native English speakers' choices in agreement with expectations, by personality dimension, with alignment.

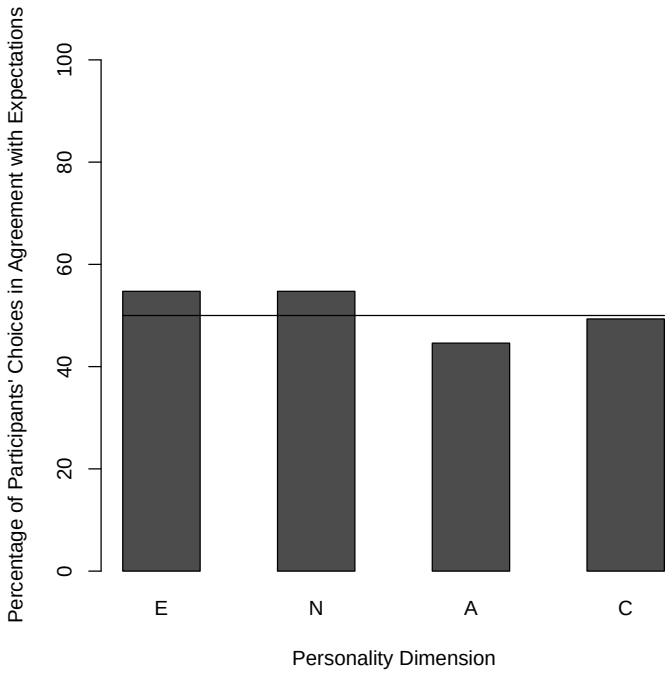

Figure 6.3: Percentage of all participants' choices in agreement with expectations, by personality dimension, with alignment.

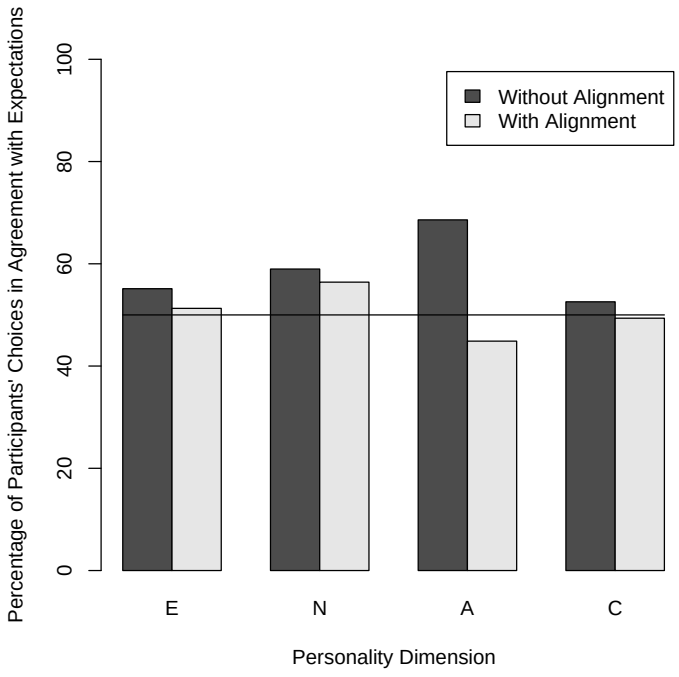

Figure 6.4: Percentage of native English speakers' choices in agreement with expectations, by personality dimension, without and with alignment.

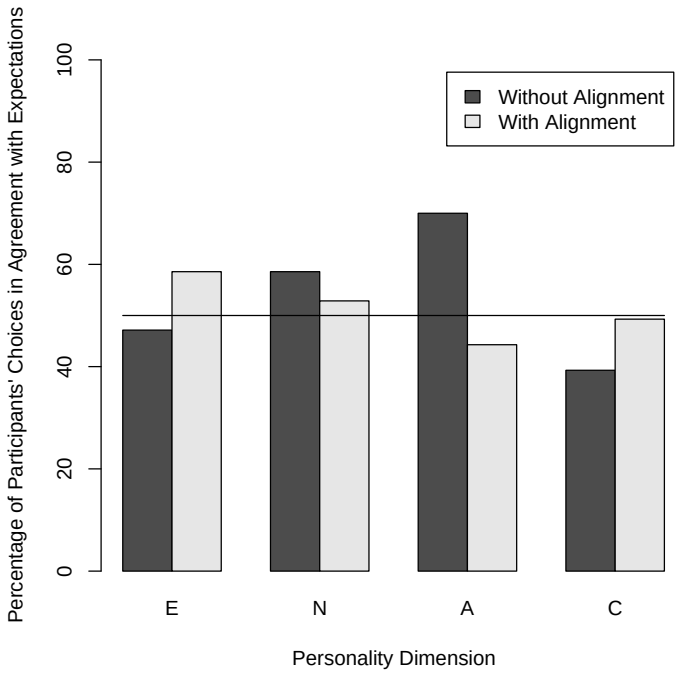

Figure 6.5: Percentage of non-native English speakers' choices in agreement with expectations, by personality dimension, without and with alignment.

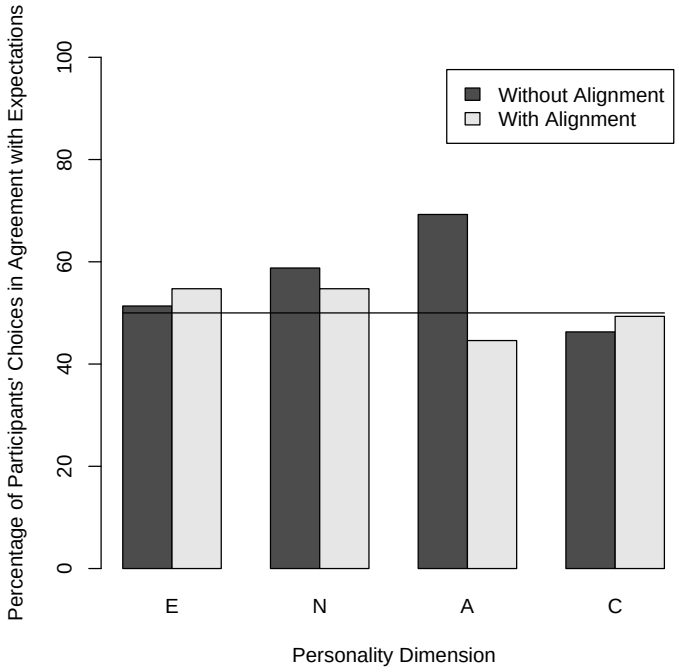

Figure 6.6: Percentage of all participants' choices in agreement with expectations, by personality dimension, without and with alignment.

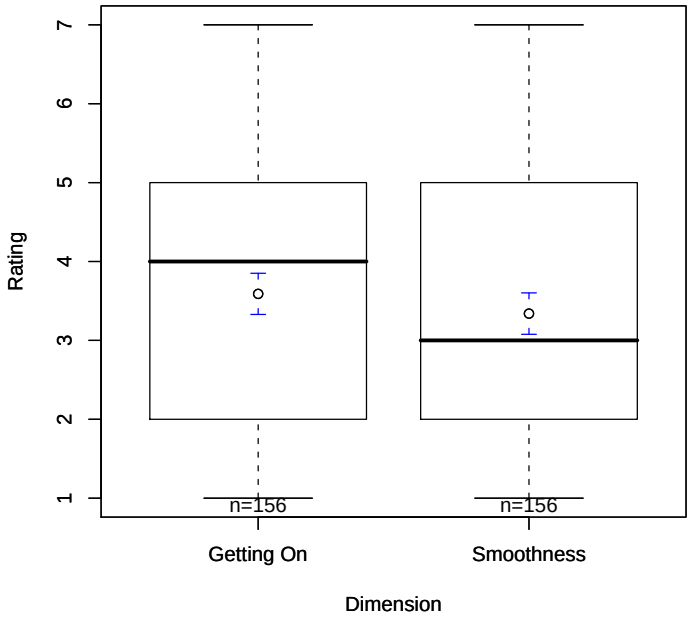

Figure 6.7: Native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went, with alignment.

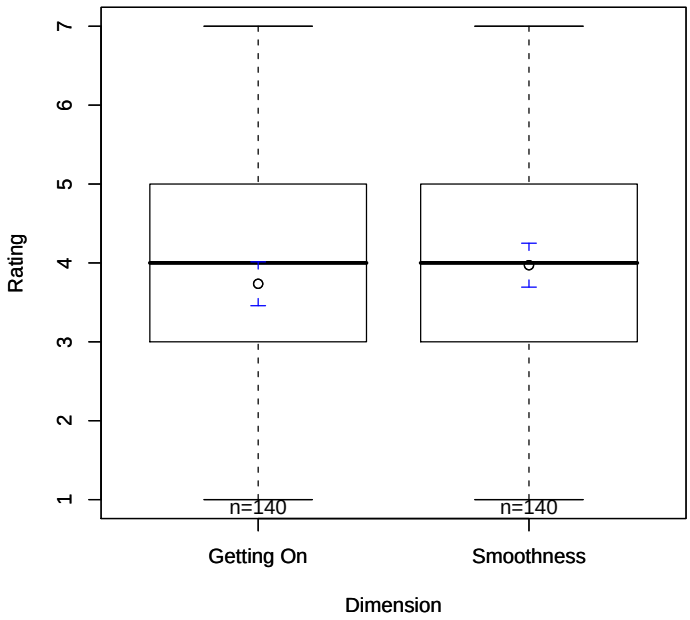

Figure 6.8: Non-native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went, with alignment.

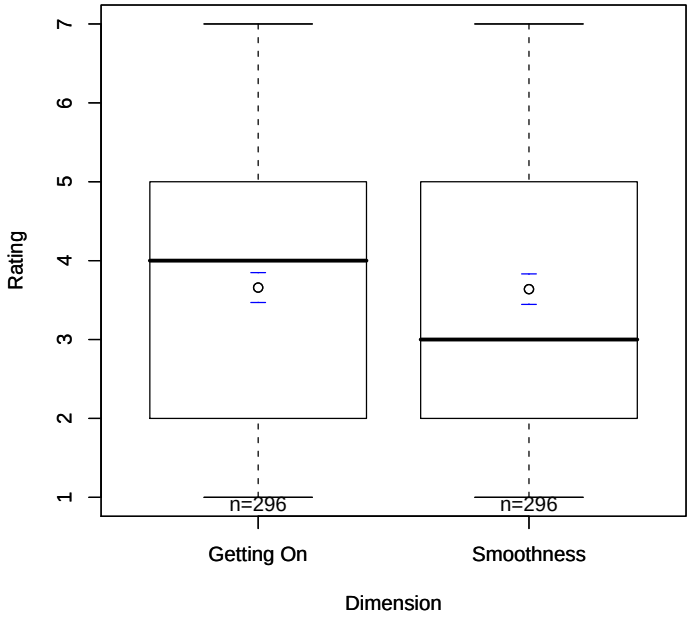

Figure 6.9: All participants' ratings of how well the characters got on with each other and how smoothly the conversation went, with alignment.

smoothness distributions without and with alignment, for the three groups of English speakers. When alignment was enabled, the mean ratings were significantly lower in all conditions ($p \leq 0.01$ for non-native speakers' smoothness ratings, $p \leq 0.001$ otherwise).

6.5 Discussion

In this chapter, we investigated the effects of adding alignment behaviour to the Affective Language Production Model. One hypothesis was that the dynamic effects of alignment would overwrite the long-term stable personality effects and thus make personality more difficult to perceive. We found that the perception of extraversion, neuroticism and conscientiousness did not change significantly, while people agreed significantly less with the model with respect to the agreeableness dimension. As the only difference between stimuli in the experiments evaluating ALPM-3 and ALPM-4 was the alignment behaviour of the second character, we conclude that alignment

Condition	n	$\bar{x}$	s	t	df	p
Native, getting on				4.669	303.431	≤ 0.001
Alignment disabled	156	4.404	1.422			
Alignment enabled	156	3.590	1.650			
Native, smoothness				4.079	309.984	≤ 0.001
Alignment disabled	156	4.109	1.672			
Alignment enabled	156	3.340	1.660			
Non-native, getting on				3.903	277.244	≤ 0.001
Alignment disabled	140	4.493	1.580			
Alignment enabled	140	3.736	1.665			
Non-native, smoothness				3.154	277.874	≤ 0.01
Alignment disabled	140	4.593	1.631			
Alignment enabled	140	3.971	1.666			
All, getting on				6.068	584.124	≤ 0.001
Alignment disabled	296	4.446	1.497			
Alignment enabled	296	3.659	1.656			
All, smoothness				5.068	589.896	≤ 0.001
Alignment disabled	296	4.338	1.667			
Alignment enabled	296	3.639	1.690			

Table 6.3: Comparison of getting on and smoothness ratings without and with alignment, for native English speakers, non-native English speakers and all speakers.

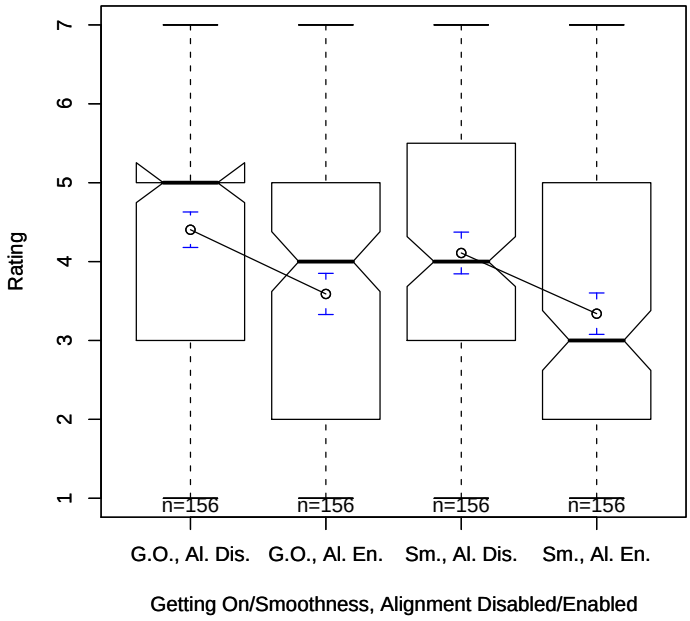

Figure 6.10: Native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went, without and with alignment.

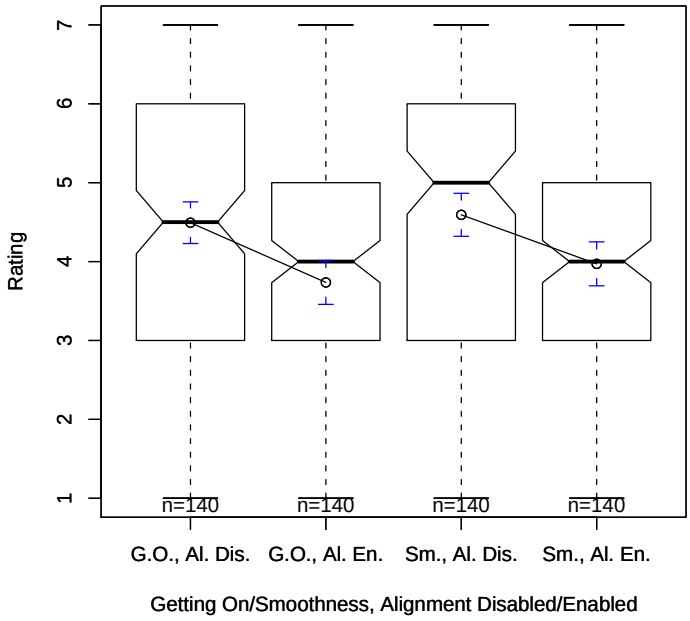

Figure 6.11: Non-native English speakers' ratings of how well the characters got on with each other and how smoothly the conversation went, without and with alignment.

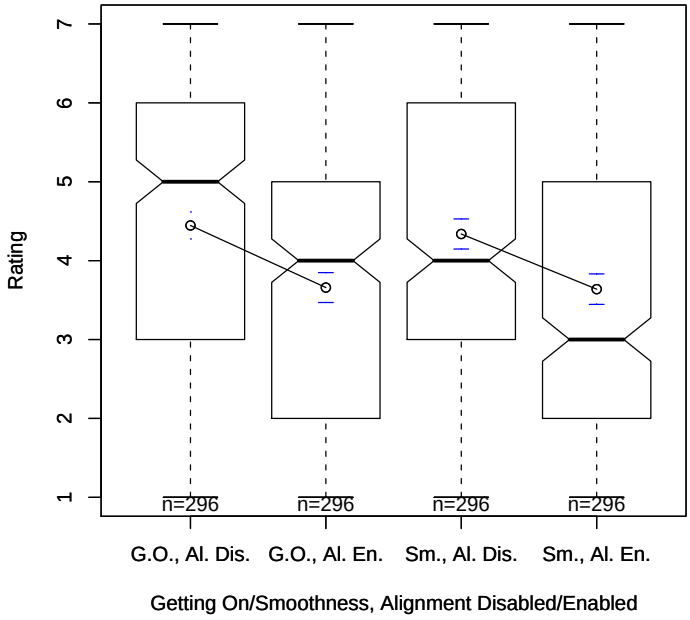

Figure 6.12: All participants' ratings of how well the characters got on with each other and how smoothly the conversation went, without and with alignment.

abolishes the accurate perception of agreeableness.

The experiment in this chapter also showed that in dialogues which exhibited alignment behaviour, the dialogue participants were perceived to be getting on with each other significantly less well, and the conversation was rated as significantly less smooth. This confirms our hypothesis that alignment has an effect on the perception of the interaction. The alignment of one of the characters had been set quite high so as to achieve an observable effect. It might be the case that too much alignment was the reason for the dialogues to be rated worse.

The cache language models are computed solely on the basis of the immediately preceding utterance. An avenue for further research would be to model a memory effect which also includes utterances that occurred earlier in the dialogue, probably with reduced weight for utterances that are less recent. It could also be studied whether to include both speakers' utterances or only those of the respective dialogue partner in this model.

Chapter 7

Conclusion

7.1 Summary

In this thesis, we explored the interaction of personality and alignment processes in natural language dialogue by developing lexically-based computational models capable of generating these variations. The models were evaluated in web-based experiments with human judges and were compared to psycholinguistic findings by way of computational simulations of priming studies.

In Chapter 3, features identified in previous studies on the relation of language and personality were combined in order to rank arbitrary utterances by personality scores. A corpus of movie review dialogues was collected. Utterances from this corpus and strategies from the Affective Language Production Model version 1 (ALPM-1) and ALPM-2 were employed to generate dialogues between computer characters with parameterisable personality. Consistent with previous findings, our web-based evaluation showed that judges were able to detect extraversion according to the model's expectations, while neuroticism was more difficult to identify. Psychoticism, which was modelled by the characters' topic selection strategy, was perceived as willingness to engage with another's opinion. The results held for both native and non-native speakers of English.

Chapter 4 contributed ALPM-3, a model capable of generating utterances from logical forms, based on mild overgeneration of paraphrases and ranking by n-gram language models of known personality within the OpenCCG framework. Optimal parameter settings for the projection of extreme personality variation were determined experimentally. In a web-based evaluation of generated dialogues between computer characters, native speakers were able to detect personality tendencies according to the model's

expectations. Agreeableness was identified most accurately, while conscientiousness was the most difficult dimension to detect. A computer character's strategy to follow their own topic agenda was perceived to be related to low agreeableness and high conscientiousness. The dialogues were judged to be reasonably naturalistic.

In Chapter 5, we introduced the concept of cache language models as a lexically-based means of capturing priming and alignment effects. A series of experiments was conducted to replicate results of psycholinguistics priming studies on the repetition of nouns and adjectives, on semantic relatedness and on prepositional object/double object priming. Fallback to simple semantic classes in situations without exact word match allowed us to model the alignment-boosting effect of semantic relatedness. While the model could not emulate human performance exactly as far as the ordering of utterances from most to least primed was concerned, the top-ranked utterance usually corresponded to what humans would have chosen. Optimal cache language model weight settings to maximise the matching effect were determined.

Finally, Chapter 6 saw the combination of ALPM-3 and cache language models into ALPM-4, a unified model of personality and alignment. Generated dialogues were evaluated in a further condition of the web-based experiment that was conducted in Chapter 4, allowing for direct comparison of the perception of dialogues with and without alignment. It was found that the introduction of alignment significantly reduced the agreement between judges and the model for the agreeableness dimension. Additionally, the computer characters were perceived to be getting on less well with each other and their dialogues were rated as less smooth.

7.2 Contributions

The main contributions of the thesis are the various revisions of the Affective Language Production Model that were implemented and employed to generate textual dialogues between computer characters whose language varies with respect to personality and alignment:

- ALPM-1 and ALPM-2: Re-combination of corpus utterances ranked by personality based on features identified in previous research.
- ALPM-3: Overgeneration of utterance paraphrases from logical forms and subsequent ranking by n-gram language models of known personality.

- ALPM-4: An extension of ALPM-3 that incorporates cache language models in order to capture alignment effects, providing a unified model for personality and alignment.

All personality-related computational resources for the models were based on projected personality, i.e., on self-assessed rather than observer ratings. Within the Critical Agent Dialogue project, the thesis contributed to the creation of the CrAg Corpus of movie review dialogues, a novel resource of personality data, and to the development of OPENCCG grammars capable of generating movie review dialogues.

The ALPMs were evaluated in web-based perception experiments, the results of which are summarised above. In preparation for the experiments, reasonable parameter settings were methodically determined for the following aspects:

- Optimal weighting given to the personality language model as opposed to the default language model in order to maximise the variety of the generated language.
- Choice between language models based on groups of weblog authors with similar personality or based on an individual's language.
- Cache language model weighting which maximises the alignment effect.

Systematic computational simulations contributed results on the adequacy of the cache language modelling approach to emulate human behaviour in psycholinguistic priming experiments.

The results of the web-based experiments provided insights into the perception of personality and alignment in generated dialogues between computer characters, some of which were unexpected.

Extraversion could be detected reliably in ALPM-2's approach of re-ranking corpus utterances. When moving to the overgeneration-based ALPM-4, extraversion became more difficult to perceive, which is probably due to lack of variation in utterance length, as language models tend to prefer shorter utterances.

ALPM-2 predicted that tough-minded characters would follow their own topic agenda instead of sticking to their interlocutor's topic. However, judges perceived this behaviour as low psychotic. In ALPM-4, following one's own agenda was hypothesised to be related to low agreeableness and low conscientiousness, but judges associated it with low agreeableness and *high* conscientiousness.

While the introduction of lexical alignment did not influence the perception of extraversion, neuroticism and conscientiousness, it abolished the accurate perception of agreeableness and had a detrimental effect on the general judgements of interaction quality. It can be concluded that excessive alignment is perceived negatively.

7.3 Future Work

In future work, the coverage of the OPENCCG grammar employed to generate the movie review dialogues could be extended in order to be able to produce more perceptible variation. In support of this, there is also a need for much larger personality-annotated corpora than have been available up to now.

In ALPM-4, the cache language model is based solely on the previous utterance of the interlocutor. It would be interesting to experiment with a larger cache that captures more of the dialogue history. Memory effects could be modelled by reducing the weight of less recent utterances.

Cache language models could be extended with further features, e.g., richer part-of-speech tags, or *supertags* (Bangalore and Joshi 1999). With CCG (and related lexicalised theories of syntax), supertags encode the syntactic category of a lexical item. For example, a verb's annotation would then already give an indication of what kind of arguments it expects. This information provides an extended domain of locality, and the model would be less dependent on exact word sequences. Techniques from supertagging have recently been integrated into OPENCCG as *hypertagging* and have been shown to improve realisation speed and quality (Espinosa et al. 2008).

More research is also required on the interaction of lexical alignment and topic shifting, which seem to be responsible for different effects on the perception of dialogues.

These enhancements are further steps on the way towards the goal of a unified model of personality and alignment processes in dialogue.

Appendix A

Experimental Instructions

In this appendix, we reproduce the instructions that were shown to participants of the web-based experiments described in Section 3.3.4.2, called 'Experiment on the Perception of Communication Styles', and Section 4.3.6.4, called 'Experiment on the Perception of Communication Styles II'.

A.1 Experiment on the Perception of Communication Styles

Instructions

Thank you for taking part in this experiment! Please read the instructions carefully before starting. Do not hesitate to contact the experimenter in case you have any questions or comments concerning the experiment.

Everyone who completes this experiment will be entered into a draw for a **£15 Amazon.co.uk gift certificate**. If you want to be entered into the draw, please make sure that you specify a valid e-mail address in the form at the bottom of this page. The address will only be used for the prize draw.

Your Task

You will be presented with a series of **eight dialogues** between two anonymous characters, in English and in textual form. The characters are discussing a movie; imagine that you are overhearing a part of their conversation.

After reading each dialogue, please **judge which of the two characters scores**

higher on each of three dimensions: Extraversion, emotional instability and tough-mindedness. You are asked to choose the higher-scoring character – sometimes it will be difficult to choose, but please go with your first impression. The dialogues are not related to each other; please rate each of them independently.

The experiment will take about **15 minutes**. Native speakers of any language are welcome to take part.

Descriptions of the Dimensions to Judge

Please base your judgements on the following descriptions (modified from Eysenck and Eysenck 1975, pp. 9–12):

Extraversion Typical extraverts are sociable, like parties, have many friends, need to have people to talk to, and do not like reading or studying by themselves. They crave excitement, take chances, often stick their neck out, act on the spur of the moment, and are generally impulsive individuals. They are fond of practical jokes, always have a ready answer, and generally like change; they are carefree, easy-going, optimistic, and like to "laugh and be merry." They prefer to keep moving and doing things, tend to be aggressive and lose their temper quickly; altogether their feelings are not kept under tight control, and they are not always reliable people.

Emotional Instability Typical emotionally unstable people are anxious, worrying individuals, moody and frequently depressed. They are likely to sleep badly, and to suffer from various psychosomatic disorders. They are overly emotional, reacting too strongly to all sorts of stimuli, and find it difficult to get back on an even keel after each emotionally arousing experience. Their strong emotional reactions interfere with their proper adjustment, making them react in irrational, sometimes rigid ways. If emotionally unstable individuals have to be described in one word, one might say that they are *worriers*; their main characteristic is a constant preoccupation with things that might go wrong, and a strong emotional reaction of anxiety to these thoughts.

Tough-Mindedness Tough-minded individuals may be described as being solitary, not caring for people; they are often troublesome, not fitting in anywhere. They may be cruel and inhumane, lacking in feeling and empathy, and altogether insensitive. They are hostile to others, even with their own kith and kin, and aggressive even

to loved ones. They have a liking for odd and unusual things, and a disregard for danger; they like to make fools of other people and to upset them.

Eysenck, H. J. and S. B. G. Eysenck (1975). *Manual of the Eysenck Personality Questionnaire (Junior & Adult)*. Sevenoaks, UK: Hodder and Stoughton Educational.

Example Dialogue

This is an example dialogue similar to those that will be presented to you during the experiment:

Character A: "A strange thing I thought was the whole sort of sub plot with Kevin Bacon and the wife."

Character B: "The ending in this was like – oh it's all over, I felt that the movie had just, I don't know."

Character A: "I can think of very few positive things to say about it really."

Character B: "So basically we have nothing good to say about it."

Character A: "I think Tim Robbins does that kind of role really well, the slightly weird slightly spaced out."

Character B: "The acting was was really good I mean these are all very talented actors."

Character A: "I do feel it's actually down to the director a lot of it, it must be cause a lot of the things that were really glaringly bad came from the production side."

Character B: "There were all these shots of the city from from above the river, I just thought that was the movie trying to to be too serious about itself, it's all these long ponderous shots."

Information about Yourself

In the form below, please enter details about yourself. **The personal data you give us is used only for scientific purposes. We will not give any of this information to anyone else, and nor will we report any information in any way that can be identified with you.**

Taking part in this experiment is entirely voluntary! Obviously, we would be grateful if you stayed the course, but of course you are at liberty to break off at any point during the experiment.

- *Age range: [(Please choose.)]
- *Gender: [(Please choose.)]
- *Hand you prefer to use for writing: [(Please choose.)]
- *Native language: _____
- *Region you grew up in: _____
- *Academic subject you study/studied, or occupation: _____
- E-mail address (for prize draw): _____
- Submit and Begin Experiment
- * denotes required field

A.2 Experiment on the Perception of Communication Styles II

Instructions

Thank you for taking part in this experiment! Please read the instructions carefully before starting. Do not hesitate to contact the experimenter in case you have any questions or comments concerning the experiment.

Everyone who completes this experiment will be entered into a draw for a **£15 Amazon.co.uk gift certificate**. If you want to be entered into the draw, please make sure that you specify a valid e-mail address in the form at the bottom of this page. The address will only be used for the prize draw.

Your Task

You will be presented with a series of **eight dialogues** between two anonymous characters, in English and in textual form. The characters are discussing a movie; imagine that you are overhearing a part of their conversation.

After reading each dialogue, please **judge which of the two characters scores higher on each of four dimensions**: Extraversion, emotional instability, agreeableness and conscientiousness. You are asked to choose the higher-scoring character – sometimes it will be difficult to choose, but please go with your first impression. You are also asked to assess (on a scale from 1 to 7) **how well the characters got on with each other** and **how smoothly the conversation went**. The dialogues are not related to each other; please rate each of them independently.

The experiment will take about **15 minutes**. Native speakers of any language are welcome to take part.

Descriptions of the Dimensions to Judge

Please base your judgements on the following descriptions (modified from Buchanan et al. 1999):

Extraversion This trait reflects preference for, and behavior in, social situations. People high in extraversion are energetic and seek out the company of others. Low scorers (introverts) tend to be more quiet and reserved.

Emotional Instability This trait reflects the tendency to experience negative thoughts and feelings. High scorers are prone to insecurity and emotional distress. Low scorers tend to be more relaxed, less emotional and less prone to distress.

Agreeableness This trait reflects how we tend to interact with others. People high in agreeableness tend to be trusting, friendly and cooperative. Low scorers tend to be more aggressive and less cooperative.

Conscientiousness This trait reflects how organised and persistent we are in pursuing our goals. High scorers are methodical, well organised and dutiful. Low scorers are less careful, less focussed and more likely to be distracted from tasks.

Buchanan, T., L. R. Goldberg, and J. A. Johnson (1999). WWW personality assessment: Evaluation of an on-line Five Factor Inventory. In *Proceedings of the 29th Annual Meeting of the Society for Computers in Psychology (SCiP-99)*, Los Angeles, CA, USA.

Example Dialogue

This is an example dialogue similar to those that will be presented to you during the experiment:

Character A: "To be honest I didn't think that the music was bad."

Character B: "I mean I didn't think that the music was great."

Character A: "I mean I didn't think that the film was bad."

Character B: "You know the movie was good."

Character A: "And you know I didn't think that the action scenes were bad."

Character B: "Yeah I mean I kind of thought that the action scenes were also pretty good."

Character A: "But to be honest I thought that there was no characterization at all."

Character B: "Well I mean I didn't think that the characterization was bad."

Information about Yourself

In the form below, please enter details about yourself. **The personal data you give us is used only for scientific purposes. We will not give any of this information to anyone else, and nor will we report any information in any way that can be identified with you.**

Taking part in this experiment is entirely voluntary! Obviously, we would be grateful if you stayed the course, but of course you are at liberty to break off at any point during the experiment.

- *Age range: [(Please choose.)]
- *Gender: [(Please choose.)]
- *Hand you prefer to use for writing: [(Please choose.)]
- *Native language: _____
- *Region you grew up in: _____
- *Academic subject you study/studied, or occupation: _____

- E-mail address (for prize draw): _____
- Submit and Begin Experiment
- * denotes required field

Appendix B

Experimental Results

B.1 Modelling Alignment with Cache Language Models

The following tables show the results of the three experiments that were described in Chapter 5.

B.1.1 Experiment 1: Repetition of Nouns and Adjectives

Cache	Target: Pre-Nominal							
LM	Prime: P-N, SN, SA		Prime: P-N, DN, SA		Prime: P-N, SN, DA		Prime: P-N, DN, DA	
Weight	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD
0.0	0.000221	0.000373	0.000221	0.000373	0.000221	0.000373	0.000221	0.000373
0.1	0.032800	0.005700	0.019800	0.003920	0.016000	0.002730	0.011100	0.002160
0.2	0.070100	0.008360	0.040900	0.006150	0.034100	0.004010	0.022800	0.003380
0.3	0.112000	0.009560	0.063800	0.007360	0.054200	0.004590	0.035500	0.004040
0.4	0.157000	0.009790	0.088100	0.007780	0.076000	0.004700	0.049000	0.004280
0.5	0.205000	0.009300	0.114000	0.007590	0.099300	0.004470	0.063200	0.004170
0.6	0.255000	0.008240	0.141000	0.006870	0.124000	0.003960	0.078100	0.003780
0.7	0.309000	0.006720	0.168000	0.005700	0.150000	0.003230	0.093600	0.003140
0.8	0.364000	0.004810	0.197000	0.004140	0.177000	0.002310	0.110000	0.002280
0.9	0.422000	0.002560	0.227000	0.002230	0.204000	0.001230	0.126000	0.001230
1.0	**0.481000**	0.000000	**0.258000**	0.000000	**0.233000**	0.000000	**0.143000**	0.000000

Table B.1: Mean geometric mean (GM) scores (with standard deviations) of pre-nominal targets following pre-nominal (P-N) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). Peak values are emphasised.

Cache	Target: Pre-Nominal							
LM	Prime: RC, SN, SA		Prime: RC, DN, SA		Prime: RC, SN, DA		Prime: RC, DN, DA	
Weight	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD
0.0	0.000221	0.000373	0.000221	0.000373	0.000221	0.000373	0.000221	0.000373
0.1	0.005500	0.001130	0.003810	0.000859	0.005030	0.001150	0.003120	0.000690
0.2	0.011000	0.001790	0.007520	0.001180	0.009890	0.001750	0.006100	0.000957
0.3	0.016700	0.002180	0.011500	0.001350	0.015100	0.002090	0.009270	0.001110
0.4	0.022800	0.002350	0.015600	0.001410	0.020500	0.002230	0.012600	0.001160
0.5	0.029200	0.002330	0.019900	0.001370	0.026200	0.002200	0.016100	0.001130
0.6	0.035700	0.002140	0.024400	0.001230	0.032000	0.002010	0.019800	0.001020
0.7	0.042500	0.001800	0.029100	0.001020	0.038100	0.001680	0.023600	0.000847
0.8	0.049400	0.001330	0.033900	0.000743	0.044300	0.001240	0.027500	0.000619
0.9	0.056600	0.000722	0.038800	0.000399	0.050700	0.000669	0.031500	0.000332
1.0	**0.063900**	0.000000	**0.043800**	0.000000	**0.057200**	0.000000	**0.035700**	0.000000

Table B.2: Mean geometric mean (GM) scores (with standard deviations) of pre-nominal targets following relative clause (RC) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). Peak values are emphasised.

Cache	Target: Relative Clause							
LM	Prime: P-N, SN, SA		Prime: P-N, DN, SA		Prime: P-N, SN, DA		Prime: P-N, DN, DA	
Weight	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD
0.0	0.000627	0.00026	0.000627	0.00026	0.000627	0.00026	0.000627	0.000260
0.1	0.016200	0.00145	0.012500	0.00108	0.012600	0.00111	0.009680	0.000821
0.2	0.024200	0.00222	0.018500	0.00164	0.019000	0.00179	0.014500	0.001310
0.3	0.030300	0.00297	0.023100	0.00220	0.023900	0.00241	0.018200	0.001800
0.4	0.034800	0.00362	0.026500	0.00271	0.027500	0.00295	0.021000	0.002210
0.5	0.037900	0.00413	0.028900	0.00312	0.030100	0.00336	0.023000	0.002530
0.6	**0.039600**	0.00448	**0.030200**	0.00339	**0.031500**	0.00364	**0.024000**	0.002750
0.7	0.039500	0.00461	0.030100	0.00350	**0.031500**	0.00373	**0.024000**	0.002830
0.8	0.037000	0.00445	0.028200	0.00338	0.029600	0.00358	0.022500	0.002720
0.9	0.030600	0.00376	0.023300	0.00286	0.024500	0.00302	0.018700	0.002300
1.0	0.000000	0.00000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000

Table B.3: Mean geometric mean (GM) scores (with standard deviations) of relative clause targets following pre-nominal (P-N) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). Peak values are emphasised.

Cache LM Weight	Target: Relative Clause							
	Prime: RC, SN, SA		Prime: RC, DN, SA		Prime: RC, SN, DA		Prime: RC, DN, DA	
	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD
0.0	0.000627	0.00026	0.000627	0.00026	0.000627	0.000260	0.000627	0.000260
0.1	0.047500	0.00470	0.031100	0.00328	0.033800	0.002580	0.022100	0.001490
0.2	0.088400	0.00623	0.056700	0.00423	0.060800	0.003260	0.039000	0.001850
0.3	0.130000	0.00684	0.082600	0.00458	0.087400	0.003390	0.055600	0.001920
0.4	0.172000	0.00687	0.109000	0.00456	0.114000	0.003240	0.072200	0.001840
0.5	0.215000	0.00645	0.136000	0.00426	0.141000	0.002910	0.089000	0.001660
0.6	0.259000	0.00567	0.163000	0.00373	0.168000	0.002470	0.106000	0.001410
0.7	0.303000	0.00460	0.191000	0.00302	0.196000	0.001940	0.123000	0.001110
0.8	0.348000	0.00328	0.219000	0.00214	0.224000	0.001340	0.140000	0.000769
0.9	0.394000	0.00173	0.247000	0.00113	0.252000	0.000693	0.158000	0.000395
1.0	**0.441000**	0.00000	**0.276000**	0.00000	**0.280000**	0.000000	**0.175000**	0.000000

Table B.4: Mean geometric mean (GM) scores (with standard deviations) of relative clause targets following relative clause (RC) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). Peak values are emphasised.

Cache	Target: Pre-Nominal							
LM	Prime: P-N, SN, SA		Prime: P-N, DN, SA		Prime: P-N, SN, DA		Prime: P-N, DN, DA	
Weight	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD
0.0	$1.56e-10$	$3.81e-10$	$1.56e-10$	$3.81e-10$	$1.56e-10$	$3.81e-10$	$1.56e-10$	$3.81e-10$
0.1	$3.80e-05$	$1.92e-05$	$8.56e-06$	$4.79e-06$	$4.42e-06$	$2.18e-06$	$1.50e-06$	$8.19e-07$
0.2	$3.56e-04$	$1.27e-04$	$7.26e-05$	$3.21e-05$	$4.10e-05$	$1.44e-05$	$1.25e-05$	$5.43e-06$
0.3	$1.50e-03$	$5.48e-04$	$2.68e-04$	$9.25e-05$	$1.62e-04$	$4.13e-05$	$4.62e-05$	$1.57e-05$
0.4	$3.67e-03$	$8.16e-04$	$6.98e-04$	$1.86e-04$	$4.44e-04$	$8.32e-05$	$1.20e-04$	$3.15e-05$
0.5	$8.50e-03$	$1.05e-03$	$1.50e-03$	$5.48e-04$	$9.37e-04$	$7.18e-05$	$2.55e-04$	$5.09e-05$
0.6	$1.67e-02$	$1.63e-03$	$2.67e-03$	$5.16e-04$	$2.00e-03$	$0.00e+00$	$4.79e-04$	$7.02e-05$
0.7	$2.93e-02$	$1.97e-03$	$4.67e-03$	$8.16e-04$	$3.33e-03$	$5.16e-04$	$8.21e-04$	$8.32e-05$
0.8	$4.83e-02$	$1.97e-03$	$7.50e-03$	$5.48e-04$	$5.50e-03$	$5.48e-04$	$1.00e-03$	$0.00e+00$
0.9	$7.50e-02$	$1.26e-03$	$1.15e-02$	$5.48e-04$	$8.50e-03$	$5.48e-04$	$2.00e-03$	$0.00e+00$
1.0	$\mathbf{1.11e-01}$	$0.00e+00$	$\mathbf{1.70e-02}$	$0.00e+00$	$\mathbf{1.30e-02}$	$0.00e+00$	$\mathbf{3.00e-03}$	$0.00e+00$

Table B.5: Mean probabilities (with standard deviations) of pre-nominal targets following pre-nominal (P-N) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). Peak values are emphasised.

Cache	Target: Pre-Nominal							
LM	Prime: RC, SN, SA		Prime: RC, DN, SA		Prime: RC, SN, DA		Prime: RC, DN, DA	
Weight	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD
0.0	$1.56e-10$	$3.81e-10$	$1.56e-10$	$3.81e-10$	$1.56e-10$	$3.81e-10$	$1.56e-10$	$3.81e-10$
0.1	$1.84e-07$	$1.01e-07$	$6.22e-08$	$3.60e-08$	$1.44e-07$	$8.93e-08$	$3.41e-08$	$1.86e-08$
0.2	$1.40e-06$	$6.54e-07$	$4.50e-07$	$1.89e-07$	$1.04e-06$	$5.29e-07$	$2.41e-07$	$1.03e-07$
0.3	$4.90e-06$	$1.88e-06$	$1.55e-06$	$5.11e-07$	$3.58e-06$	$1.47e-06$	$8.26e-07$	$2.82e-07$
0.4	$1.22e-05$	$3.74e-06$	$3.87e-06$	$1.00e-06$	$8.87e-06$	$2.89e-06$	$2.06e-06$	$5.56e-07$
0.5	$2.52e-05$	$6.07e-06$	$7.99e-06$	$1.59e-06$	$1.82e-05$	$4.62e-06$	$4.26e-06$	$8.84e-07$
0.6	$4.60e-05$	$8.34e-06$	$1.46e-05$	$2.19e-06$	$3.32e-05$	$6.28e-06$	$7.82e-06$	$1.21e-06$
0.7	$7.70e-05$	$9.90e-06$	$2.46e-05$	$2.56e-06$	$5.55e-05$	$7.41e-06$	$1.32e-05$	$1.43e-06$
0.8	$1.21e-04$	$9.89e-06$	$3.89e-05$	$2.54e-06$	$8.71e-05$	$7.34e-06$	$2.08e-05$	$1.42e-06$
0.9	$1.81e-04$	$7.14e-06$	$5.84e-05$	$1.81e-06$	$1.30e-04$	$5.24e-06$	$3.14e-05$	$9.91e-07$
1.0	$\mathbf{2.61e-04}$	$0.00e+00$	$\mathbf{8.43e-05}$	$0.00e+00$	$\mathbf{1.87e-04}$	$0.00e+00$	$\mathbf{4.54e-05}$	$0.00e+00$

Table B.6: Mean probabilities (with standard deviations) of pre-nominal targets following relative clause (RC) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). Peak values are emphasised.

Cache	Target: Relative Clause							
LM	Prime: P-N, SN, SA		Prime: P-N, DN, SA		Prime: P-N, SN, DA		Prime: P-N, DN, DA	
Weight	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD
0.0	$3.15e-16$	$6.12e-16$	$3.15e-16$	$6.12e-16$	$3.15e-16$	$6.12e-16$	$3.15e-16$	$6.12e-16$
0.1	$1.19e-09$	$5.18e-10$	$3.18e-10$	$1.33e-10$	$3.36e-10$	$1.41e-10$	$8.98e-11$	$3.61e-11$
0.2	$8.88e-09$	$3.64e-09$	$2.32e-09$	$9.27e-10$	$2.64e-09$	$1.09e-09$	$6.89e-10$	$2.76e-10$
0.3	$2.73e-08$	$1.16e-08$	$7.07e-09$	$2.94e-09$	$8.36e-09$	$3.62e-09$	$2.16e-09$	$9.17e-10$
0.4	$5.53e-08$	$2.45e-08$	$1.43e-08$	$6.25e-09$	$1.73e-08$	$7.86e-09$	$4.46e-09$	$2.00e-09$
0.5	$8.58e-08$	$3.97e-08$	$2.21e-08$	$1.01e-08$	$2.72e-08$	$1.29e-08$	$7.01e-09$	$3.30e-09$
0.6	$\mathbf{1.07e-07}$	$5.13e-08$	$\mathbf{2.74e-08}$	$1.31e-08$	$3.43e-08$	$1.68e-08$	$8.82e-09$	$4.30e-09$
0.7	$1.06e-07$	$5.26e-08$	$2.73e-08$	$1.35e-08$	$\mathbf{3.44e-08}$	$1.73e-08$	$\mathbf{8.84e-09}$	$4.44e-09$
0.8	$7.76e-08$	$3.95e-08$	$1.99e-08$	$1.01e-08$	$2.53e-08$	$1.30e-08$	$6.49e-09$	$3.33e-09$
0.9	$3.02e-08$	$1.57e-08$	$7.73e-09$	$4.02e-09$	$9.92e-09$	$5.20e-09$	$2.54e-09$	$1.33e-09$
1.0	$0.00e+00$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$0.00e+00$

Table B.7: Mean probabilities (with standard deviations) of relative clause targets following pre-nominal (P-N) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). Peak values are emphasised.

Cache	Target: Relative Clause							
LM	Prime: RC, SN, SA		Prime: RC, DN, SA		Prime: RC, SN, DA		Prime: RC, DN, DA	
Weight	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD
0.0	$3.15e-16$	$6.12e-16$	$3.15e-16$	$6.12e-16$	$3.15e-16$	$6.12e-16$	$3.15e-16$	$6.12e-16$
0.1	$2.60e-07$	$1.08e-07$	$3.17e-08$	$1.35e-08$	$4.63e-08$	$1.67e-08$	$5.52e-09$	$1.78e-09$
0.2	$5.60e-06$	$1.69e-06$	$6.11e-07$	$1.89e-07$	$8.49e-07$	$2.21e-07$	$9.15e-08$	$2.13e-08$
0.3	$3.77e-05$	$8.77e-06$	$3.93e-06$	$9.40e-07$	$5.17e-06$	$9.92e-07$	$5.34e-07$	$9.21e-08$
0.4	$1.53e-04$	$2.77e-05$	$1.55e-05$	$2.90e-06$	$1.95e-05$	$2.75e-06$	$1.97e-06$	$2.52e-07$
0.5	$4.64e-04$	$6.41e-05$	$4.64e-05$	$6.65e-06$	$5.60e-05$	$5.76e-06$	$5.59e-06$	$5.22e-07$
0.6	$9.89e-04$	$2.65e-05$	$1.16e-04$	$1.24e-05$	$1.35e-04$	$1.01e-05$	$1.33e-05$	$8.82e-07$
0.7	$2.83e-03$	$4.08e-04$	$2.52e-04$	$1.89e-05$	$2.88e-04$	$1.45e-05$	$2.82e-05$	$1.27e-06$
0.8	$5.00e-03$	$0.00e+00$	$5.01e-04$	$2.38e-05$	$5.58e-04$	$1.69e-05$	$5.44e-05$	$1.48e-06$
0.9	$9.83e-03$	$4.08e-04$	$9.24e-04$	$2.09e-05$	$9.98e-04$	$4.02e-06$	$9.78e-05$	$1.23e-06$
1.0	$\mathbf{1.70e-02}$	$0.00e+00$	$\mathbf{2.00e-03}$	$0.00e+00$	$\mathbf{2.00e-03}$	$0.00e+00$	$\mathbf{1.66e-04}$	$0.00e+00$

Table B.8: Mean probabilities (with standard deviations) of relative clause targets following relative clause (RC) primes with same noun (SN), different noun (DN), same adjective (SA) or different adjective (DA). Peak values are emphasised.

B.1.2 Experiment 2: Semantic Relatedness

Cache	Target: Pre-Nominal											
LM	Prime: P-N, SN		Prime: P-N, SRN		Prime: P-N, SUN		Prime: RC, SN		Prime: RC, SRN		Prime: RC, SUN	
Weight	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD
0.0	$8.26e-05$	$6.31e-05$	$8.26e-05$	$6.31e-05$	$8.26e-05$	$6.31e-05$	$8.26e-05$	$6.31e-05$	$8.26e-05$	$6.31e-05$	$8.26e-05$	$6.31e-05$
0.1	$3.12e-02$	$4.33e-03$	$1.86e-02$	$3.14e-03$	$1.47e-03$	$8.29e-04$	$5.04e-03$	$9.49e-04$	$3.32e-03$	$5.97e-04$	$6.12e-04$	$3.42e-04$
0.2	$6.77e-02$	$6.10e-03$	$3.90e-02$	$4.70e-03$	$2.35e-03$	$1.27e-03$	$1.03e-02$	$1.45e-03$	$6.84e-03$	$8.84e-04$	$9.64e-04$	$5.18e-04$
0.3	$1.09e-01$	$6.81e-03$	$6.15e-02$	$5.45e-03$	$3.09e-03$	$1.62e-03$	$1.60e-02$	$1.70e-03$	$1.07e-02$	$1.03e-03$	$1.26e-03$	$6.57e-04$
0.4	$1.54e-01$	$6.85e-03$	$8.57e-02$	$5.64e-03$	$3.71e-03$	$1.91e-03$	$2.20e-02$	$1.79e-03$	$1.48e-02$	$1.07e-03$	$1.50e-03$	$7.71e-04$
0.5	$2.02e-01$	$6.42e-03$	$1.11e-01$	$5.40e-03$	$4.21e-03$	$2.14e-03$	$2.83e-02$	$1.74e-03$	$1.92e-02$	$1.03e-03$	$1.69e-03$	$8.61e-04$
0.6	$2.53e-01$	$5.62e-03$	$1.38e-01$	$4.81e-03$	$4.56e-03$	$2.29e-03$	$3.50e-02$	$1.57e-03$	$2.37e-02$	$9.18e-04$	$1.83e-03$	$9.23e-04$
0.7	$3.07e-01$	$4.55e-03$	$1.67e-01$	$3.94e-03$	$\mathbf{4.73e-03}$	$2.36e-03$	$4.18e-02$	$1.30e-03$	$2.85e-02$	$7.54e-04$	$\mathbf{1.90e-03}$	$9.48e-04$
0.8	$3.63e-01$	$3.23e-03$	$1.96e-01$	$2.83e-03$	$4.65e-03$	$2.31e-03$	$4.90e-02$	$9.35e-04$	$3.34e-02$	$5.46e-04$	$1.86e-03$	$9.23e-04$
0.9	$4.21e-01$	$1.71e-03$	$2.26e-01$	$1.51e-03$	$4.09e-03$	$2.03e-03$	$5.63e-02$	$5.07e-04$	$3.86e-02$	$2.91e-04$	$1.63e-03$	$8.07e-04$
1.0	$\mathbf{4.81e-01}$	$0.00e+00$	$\mathbf{2.58e-01}$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$\mathbf{6.39e-02}$	$0.00e+00$	$\mathbf{4.38e-02}$	$0.00e+00$	$0.00e+00$	$0.00e+00$

Table B.9: Mean geometric mean (GM) scores (with standard deviations) of pre-nominal targets following pre-nominal (P-N) or relative clause (RC) primes with same noun (SN), semantically related noun (SRN) or semantically unrelated noun (SUN). Peak values are emphasised.

Cache	Target: Relative Clause											
LM	Prime: P-N, SN		Prime: P-N, SRN		Prime: P-N, SUN		Prime: RC, SN		Prime: RC, SRN		Prime: RC, SUN	
Weight	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD
0.0	0.0012	0.0005	0.0012	0.0005	0.0012	0.0005	0.0012	0.0005	0.0012	0.0005	0.0012	0.0005
0.1	0.0207	0.0027	0.0159	0.0022	0.0075	0.0036	0.0553	0.0051	0.0374	0.0034	0.0110	0.0051
0.2	0.0304	0.0041	0.0232	0.0032	0.0094	0.0045	0.0980	0.0076	0.0643	0.0047	0.0157	0.0071
0.3	0.0375	0.0051	0.0286	0.0040	0.0104	0.0050	0.1401	0.0090	0.0905	0.0054	0.0196	0.0087
0.4	0.0427	0.0060	0.0326	0.0046	**0.0108**	0.0052	0.1822	0.0097	0.1167	0.0057	0.0229	0.0100
0.5	0.0463	0.0066	0.0353	0.0051	0.0108	0.0052	0.2245	0.0096	0.1429	0.0056	0.0257	0.0111
0.6	**0.0480**	0.0070	**0.0366**	0.0054	0.0103	0.0050	0.2671	0.0089	0.1692	0.0052	0.0279	0.0119
0.7	0.0477	0.0071	0.0363	0.0054	0.0094	0.0045	0.3100	0.0076	0.1957	0.0044	0.0294	0.0124
0.8	0.0446	0.0067	0.0340	0.0051	0.0079	0.0038	0.3532	0.0057	0.2223	0.0033	**0.0299**	0.0125
0.9	0.0367	0.0056	0.0280	0.0043	0.0055	0.0027	0.3968	0.0031	0.2491	0.0018	0.0284	0.0117
1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.4406**	0.0000	**0.2761**	0.0000	0.0000	0.0000

Table B.10: Mean geometric mean (GM) scores (with standard deviations) of relative clause targets following pre-nominal (P-N) or relative clause (RC) primes with same noun (SN), semantically related noun (SRN) or semantically unrelated noun (SUN). Peak values are emphasised.

Cache	Target: Pre-Nominal											
LM	Prime: P-N, SN		Prime: P-N, SRN		Prime: P-N, SUN		Prime: RC, SN		Prime: RC, SRN		Prime: RC, SUN	
Weight	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD
0.0	$1.58e-12$	$2.98e-12$	$1.58e-12$	$2.98e-12$	$1.58e-12$	$2.98e-12$	$1.58e-12$	$2.98e-12$	$1.58e-12$	$2.98e-12$	$1.58e-12$	$2.98e-12$
0.1	$3.19e-05$	$1.28e-05$	$6.89e-06$	$3.28e-06$	$5.89e-09$	$8.11e-09$	$1.39e-07$	$7.15e-08$	$3.95e-08$	$1.95e-08$	$4.19e-10$	$5.69e-10$
0.2	$3.16e-04$	$8.46e-05$	$6.17e-05$	$2.17e-05$	$2.28e-08$	$2.87e-08$	$1.14e-06$	$4.58e-07$	$3.34e-07$	$1.24e-07$	$1.56e-09$	$1.98e-09$
0.3	$1.17e-03$	$4.08e-04$	$2.38e-04$	$6.23e-05$	$5.00e-08$	$5.91e-08$	$4.18e-06$	$1.30e-06$	$1.25e-06$	$3.55e-07$	$3.37e-09$	$4.02e-09$
0.4	$3.50e-03$	$5.48e-04$	$6.37e-04$	$1.25e-04$	$8.49e-08$	$9.56e-08$	$1.08e-05$	$2.61e-06$	$3.28e-06$	$7.04e-07$	$5.63e-09$	$6.40e-09$
0.5	$8.33e-03$	$8.16e-04$	$1.33e-03$	$5.16e-04$	$1.22e-07$	$1.33e-07$	$2.30e-05$	$4.17e-06$	$7.08e-06$	$1.14e-06$	$8.00e-09$	$8.75e-09$
0.6	$1.63e-02$	$1.21e-03$	$2.67e-03$	$5.16e-04$	$1.54e-07$	$1.63e-07$	$4.29e-05$	$5.77e-06$	$1.34e-05$	$1.57e-06$	$1.00e-08$	$1.06e-08$
0.7	$2.90e-02$	$1.55e-03$	$4.50e-03$	$5.48e-04$	$\mathbf{1.71e-07}$	$1.77e-07$	$7.34e-05$	$6.82e-06$	$2.32e-05$	$1.83e-06$	$\mathbf{1.10e-08}$	$1.14e-08$
0.8	$4.77e-02$	$1.21e-03$	$7.50e-03$	$5.48e-04$	$1.61e-07$	$1.64e-07$	$1.17e-04$	$6.56e-06$	$3.74e-05$	$1.80e-06$	$1.03e-08$	$1.05e-08$
0.9	$7.47e-02$	$8.16e-04$	$1.15e-02$	$5.48e-04$	$1.10e-07$	$1.10e-07$	$1.78e-04$	$4.85e-06$	$5.74e-05$	$1.29e-06$	$6.96e-09$	$7.01e-09$
1.0	$\mathbf{1.11e-01}$	$0.00e+00$	$\mathbf{1.70e-02}$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$\mathbf{2.61e-04}$	$0.00e+00$	$\mathbf{8.43e-05}$	$0.00e+00$	$0.00e+00$	$0.00e+00$

Table B.11: Mean probabilities (with standard deviations) of pre-nominal targets following pre-nominal (P-N) or relative clause (RC) primes with same noun (SN), semantically related noun (SRN) or semantically unrelated noun (SUN). Peak values are emphasised.

Cache	Target: Relative Clause											
LM	Prime: P-N, SN		Prime: P-N, SRN		Prime: P-N, SUN		Prime: RC, SN		Prime: RC, SRN		Prime: RC, SUN	
Weight	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD
0.0	$5.15e-15$	$4.98e-15$	$5.15e-15$	$4.98e-15$	$5.15e-15$	$4.98e-15$	$5.15e-15$	$4.98e-15$	$5.15e-15$	$4.98e-15$	$5.15e-15$	$4.98e-15$
0.1	$4.38e-09$	$2.76e-09$	$1.18e-09$	$7.68e-10$	$7.22e-11$	$9.22e-11$	$5.60e-07$	$2.97e-07$	$7.81e-08$	$3.59e-08$	$4.38e-10$	$4.79e-10$
0.2	$3.00e-08$	$2.00e-08$	$7.84e-09$	$5.31e-09$	$2.23e-10$	$3.02e-10$	$9.54e-06$	$4.32e-06$	$1.14e-06$	$4.44e-07$	$2.56e-09$	$2.71e-09$
0.3	$8.66e-08$	$6.08e-08$	$2.25e-08$	$1.59e-08$	$3.76e-10$	$5.33e-10$	$5.59e-05$	$2.09e-05$	$6.27e-06$	$2.05e-06$	$7.54e-09$	$7.83e-09$
0.4	$1.68e-07$	$1.23e-07$	$4.35e-08$	$3.21e-08$	$4.71e-10$	$6.92e-10$	$2.06e-04$	$6.22e-05$	$2.21e-05$	$5.90e-06$	$1.61e-08$	$1.65e-08$
0.5	$2.52e-07$	$1.90e-07$	$6.48e-08$	$4.92e-08$	$\mathbf{4.72e-10}$	$7.12e-10$	$5.79e-04$	$1.39e-04$	$6.03e-05$	$1.29e-05$	$2.80e-08$	$2.82e-08$
0.6	$\mathbf{3.05e-07}$	$2.38e-07$	$\mathbf{7.86e-08}$	$6.14e-08$	$3.81e-10$	$5.87e-10$	$1.17e-03$	$4.08e-04$	$1.40e-04$	$2.32e-05$	$4.15e-08$	$4.15e-08$
0.7	$2.97e-07$	$2.36e-07$	$7.63e-08$	$6.08e-08$	$2.39e-10$	$3.75e-10$	$3.17e-03$	$4.08e-04$	$2.88e-04$	$3.47e-05$	$5.31e-08$	$5.26e-08$
0.8	$2.13e-07$	$1.73e-07$	$5.47e-08$	$4.46e-08$	$1.00e-10$	$1.60e-10$	$5.17e-03$	$4.08e-04$	$5.44e-04$	$4.21e-05$	$\mathbf{5.67e-08}$	$5.61e-08$
0.9	$8.15e-08$	$6.73e-08$	$2.09e-08$	$1.73e-08$	$1.70e-11$	$2.76e-11$	$1.02e-02$	$4.08e-04$	$9.55e-04$	$2.39e-05$	$4.32e-08$	$4.26e-08$
1.0	$0.00e+00$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$0.00e+00$	$\mathbf{1.70e-02}$	$0.00e+00$	$\mathbf{2.00e-03}$	$0.00e+00$	$0.00e+00$	$0.00e+00$

Table B.12: Mean probabilities (with standard deviations) of relative clause targets following pre-nominal (P-N) or relative clause (RC) primes with same noun (SN), semantically related noun (SRN) or semantically unrelated noun (SUN). Peak values are emphasised.

B.1.3 Experiment 3: Prepositional Object/Double Object Priming

Cache	Target: Prepositional Object							
LM	Prime: PO, SV		Prime: DO, SV		Prime: PO, DV		Prime: DO, DV	
Weight	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD
0.0	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
0.1	0.0136	0.0062	0.0115	0.0053	0.0111	0.0050	0.0094	0.0044
0.2	0.0232	0.0111	0.0184	0.0088	0.0186	0.0087	0.0149	0.0072
0.3	0.0321	0.0161	0.0243	0.0120	0.0256	0.0126	0.0196	0.0098
0.4	0.0406	0.0214	0.0295	0.0150	0.0323	0.0166	0.0238	0.0123
0.5	0.0487	0.0269	0.0339	0.0177	0.0386	0.0208	0.0273	0.0145
0.6	0.0562	0.0327	0.0375	0.0201	0.0445	0.0252	0.0302	0.0165
0.7	0.0631	0.0388	0.0401	0.0221	0.0499	0.0298	0.0322	0.0181
0.8	0.0690	0.0453	**0.0411**	0.0235	0.0545	0.0347	**0.0331**	0.0194
0.9	**0.0729**	0.0525	0.0392	0.0239	**0.0575**	0.0403	0.0316	0.0197
1.0	0.0320	0.0784	0.0000	0.0000	0.0247	0.0606	0.0000	0.0000

Table B.13: Mean geometric mean (GM) scores (with standard deviations) of prepositional object targets following prepositional object (PO) or double object (DO) primes with same verb (SV) or different verb (DV). Peak values are emphasised.

Cache	Target: Double Object							
LM	Prime: PO, SV		Prime: DO, SV		Prime: PO, DV		Prime: DO, DV	
Weight	Mean GM	SD	Mean GM	SD	Mean GM	SD	Mean GM	SD
0.0	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
0.1	0.0085	0.0039	0.0100	0.0048	0.0070	0.0033	0.0078	0.0037
0.2	0.0153	0.0076	0.0182	0.0092	0.0126	0.0062	0.0140	0.0071
0.3	0.0220	0.0114	0.0262	0.0140	0.0180	0.0093	0.0200	0.0107
0.4	0.0284	0.0154	0.0339	0.0190	0.0232	0.0125	0.0259	0.0145
0.5	0.0347	0.0197	0.0414	0.0243	0.0282	0.0159	0.0315	0.0184
0.6	0.0406	0.0240	0.0485	0.0297	0.0329	0.0195	0.0369	0.0226
0.7	0.0460	0.0287	0.0550	0.0355	0.0373	0.0232	0.0418	0.0270
0.8	0.0506	0.0335	0.0606	0.0416	0.0410	0.0272	0.0460	0.0316
0.9	**0.0535**	0.0389	**0.0642**	0.0484	**0.0434**	0.0315	**0.0488**	0.0367
1.0	0.0234	0.0573	0.0288	0.0706	0.0189	0.0464	0.0219	0.0536

Table B.14: Mean geometric mean (GM) scores (with standard deviations) of double object targets following prepositional object (PO) or double object (DO) primes with same verb (SV) or different verb (DV). Peak values are emphasised.

Cache	Target: Prepositional Object							
LM	Prime: PO, SV		Prime: DO, SV		Prime: PO, DV		Prime: DO, DV	
Weight	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD
0.0	$2.08e-23$	$5.10e-23$	$2.08e-23$	$5.10e-23$	$2.08e-23$	$5.10e-23$	$2.08e-23$	$5.10e-23$
0.1	$1.24e-14$	$1.97e-14$	$4.28e-15$	$9.11e-15$	$2.33e-15$	$3.99e-15$	$8.52e-16$	$1.78e-15$
0.2	$1.07e-12$	$1.71e-12$	$1.89e-13$	$3.76e-13$	$1.67e-13$	$2.54e-13$	$3.52e-14$	$6.72e-14$
0.3	$1.95e-11$	$3.67e-11$	$1.86e-12$	$3.44e-12$	$2.78e-12$	$4.89e-12$	$3.39e-13$	$5.96e-13$
0.4	$1.74e-10$	$3.64e-10$	$9.16e-12$	$1.58e-11$	$2.37e-11$	$4.75e-11$	$1.67e-12$	$2.72e-12$
0.5	$1.01e-09$	$2.25e-09$	$2.99e-11$	$4.83e-11$	$1.33e-10$	$2.90e-10$	$5.48e-12$	$8.52e-12$
0.6	$4.38e-09$	$1.02e-08$	$7.27e-11$	$1.12e-10$	$5.71e-10$	$1.30e-09$	$1.36e-11$	$2.09e-11$
0.7	$1.55e-08$	$3.69e-08$	$1.38e-10$	$2.13e-10$	$2.00e-09$	$4.71e-09$	$2.65e-11$	$4.25e-11$
0.8	$4.71e-08$	$1.14e-07$	$\mathbf{2.03e-10}$	$3.38e-10$	$6.04e-09$	$1.45e-08$	$4.02e-11$	$7.13e-11$
0.9	$1.26e-07$	$3.08e-07$	$2.01e-10$	$3.92e-10$	$1.61e-08$	$3.92e-08$	$\mathbf{4.14e-11}$	$8.52e-11$
1.0	$\mathbf{3.08e-07}$	$7.55e-07$	$0.00e+00$	$0.00e+00$	$\mathbf{3.92e-08}$	$9.59e-08$	$0.00e+00$	$0.00e+00$

Table B.15: Mean probabilities (with standard deviations) of prepositional object targets following prepositional object (PO) or double object (DO) primes with same verb (SV) or different verb (DV). Peak values are emphasised.

Cache	Target: Double Object							
LM	Prime: PO, SV		Prime: DO, SV		Prime: PO, DV		Prime: DO, DV	
Weight	Mean P	SD	Mean P	SD	Mean P	SD	Mean P	SD
0.0	$8.73e-23$	$2.14e-22$	$8.73e-23$	$2.14e-22$	$8.73e-23$	$2.14e-22$	$8.73e-23$	$2.14e-22$
0.1	$2.42e-14$	$3.95e-14$	$7.95e-14$	$1.24e-13$	$6.72e-15$	$1.13e-14$	$1.37e-14$	$2.17e-14$
0.2	$1.77e-12$	$2.69e-12$	$6.47e-12$	$1.03e-11$	$4.42e-13$	$6.68e-13$	$1.01e-12$	$1.59e-12$
0.3	$2.71e-11$	$4.60e-11$	$1.05e-10$	$1.90e-10$	$6.47e-12$	$1.09e-11$	$1.58e-11$	$2.85e-11$
0.4	$2.03e-10$	$3.85e-10$	$8.15e-10$	$1.63e-09$	$4.78e-11$	$8.99e-11$	$1.21e-10$	$2.41e-10$
0.5	$1.02e-09$	$2.08e-09$	$4.16e-09$	$8.90e-09$	$2.35e-10$	$4.79e-10$	$6.13e-10$	$1.31e-09$
0.6	$3.86e-09$	$8.39e-09$	$1.61e-08$	$3.61e-08$	$8.89e-10$	$1.93e-09$	$2.36e-09$	$5.28e-09$
0.7	$1.21e-08$	$2.75e-08$	$5.13e-08$	$1.19e-07$	$2.78e-09$	$6.29e-09$	$7.48e-09$	$1.73e-08$
0.8	$3.31e-08$	$7.74e-08$	$1.41e-07$	$3.34e-07$	$7.55e-09$	$1.77e-08$	$2.04e-08$	$4.83e-08$
0.9	$8.04e-08$	$1.93e-07$	$3.46e-07$	$8.35e-07$	$1.83e-08$	$4.39e-08$	$5.01e-08$	$1.21e-07$
1.0	$\mathbf{1.78e-07}$	$4.37e-07$	$\mathbf{7.75e-07}$	$1.90e-06$	$\mathbf{4.08e-08}$	$1.00e-07$	$\mathbf{1.12e-07}$	$2.75e-07$

Table B.16: Mean probabilities (with standard deviations) of double object targets following prepositional object (PO) or double object (DO) primes with same verb (SV) or different verb (DV). Peak values are emphasised.

Bibliography

André, E., T. Rist, S. van Mulken, M. Klesen, and S. Baldes (2000). The automated design of believable dialogues for animated presentation teams. See Cassell et al. (2000), Chapter 8, pp. 220–255.

Argamon, S., S. Dhawle, M. Koppel, and J. W. Pennebaker (2005). Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America (Interface/CSNA-05)*, St. Louis, MO, USA.

Bangalore, S. and A. K. Joshi (1999). Supertagging: An approach to almost parsing. *Computational Linguistics 25*(2), 237–265.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology 18*(3), 355–387.

Bock, K. (1989). Closed-class immanence in sentence production. *Cognition 31*(2), 163–186.

Bock, K. and H. Loebell (1990). Framing sentences. *Cognition 35*(1), 1–39.

Branigan, H. P., M. J. Pickering, and A. A. Cleland (2000). Syntactic co-ordination in dialogue. *Cognition 75*(2), B13–B25.

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the 1996 International Symposium on Spoken Dialogue (ISSD-96) at the 4th International Conference on Spoken Language Processing (ICSLP-96)*, Philadelphia, PA, USA, pp. 41–44.

Brockmann, C., A. Isard, J. Oberlander, and M. White (2005). Modelling alignment for affective dialogue. In *Proceedings of the Workshop on Adapting the Interaction Style to Affective Factors at the 10th International Conference on User Modeling (UM-05)*, Edinburgh, UK.

Buchanan, T., L. R. Goldberg, and J. A. Johnson (1999). WWW personality assessment: Evaluation of an on-line Five Factor Inventory. In *Proceedings of the 29th Annual Meeting of the Society for Computers in Psychology (SCiP-99)*, Los Angeles, CA, USA.

Carenini, G. and J. D. Moore (2000). A strategy for generating evaluative arguments. In *Proceedings of the 1st International Natural Language Generation Conference (INLG-00)*, Mitzpe Ramon, Israel, pp. 47–54.

Cassell, J., J. Sullivan, S. Prevost, and E. Churchill (Eds.) (2000). *Embodied Conversational Agents*. Cambridge, MA, USA: MIT Press.

Cleland, A. A. and M. J. Pickering (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language 49*(2), 214–230.

Costa, Jr., P. T. and R. R. McCrae (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Odessa, FL, USA: Psychological Assessment Resources.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7*, 1–30.

Dewaele, J.-M. and A. Furnham (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning 49*(3), 509–544.

Dubey, A., P. Sturt, and F. Keller (2005). Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-05)*, Vancouver, BC, Canada, pp. 827–834.

Edwards, W. and F. H. Barron (1994). SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes 60*(3), 306–325.

Espinosa, D., M. White, and D. Mehay (2008). Hypertagging: Supertagging for surface realization with CCG. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, Columbus, OH, USA, pp. 183–191.

Eysenck, H. J. and S. B. G. Eysenck (1975). *Manual of the Eysenck Personality Questionnaire (Junior & Adult)*. Sevenoaks, UK: Hodder and Stoughton Educational.

Eysenck, S. B. G., H. J. Eysenck, and P. Barrett (1985). A revised version of the psychoticism scale. *Personality and Individual Differences 6*(1), 21–29.

Foster, M. E. and M. White (2004). Techniques for text planning with XSLT. In *Proceedings of the 4th Workshop on NLP and XML (NLPXML-04) at the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, pp. 1–8.

Foster, M. E. and M. White (2005). Assessing the impact of adaptive generation in the COMIC multimodal dialogue system. In *Proceedings of the 4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems (KRPDS-05) at the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, Edinburgh, UK.

Garrod, S. and M. J. Pickering (2004). Why is conversation so easy? *Trends in Cognitive Sciences 8*(1), 8–11.

Gill, A. J. (2004). *Personality and Language: The projection and perception of personality in computer-mediated communication*. Ph. D. thesis, University of Edinburgh, UK.

Gill, A. J., A. J. Harrison, and J. Oberlander (2004). Interpersonality: Individual differences and interpersonal priming. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society (CogSci2004)*, Chicago, IL, USA, pp. 464–469.

Gill, A. J. and J. Oberlander (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society (CogSci2002)*, Fairfax, VA, USA, pp. 363–368.

Gill, A. J., J. Oberlander, and E. Austin (2006). Rating e-mail personality at zero acquaintance. *Personality and Individual Differences 40*(3), 497–507.

Godfrey, J. J., E. C. Holliman, and J. McDaniel (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP-92)*, San Francisco, CA, USA, pp. 517–520.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment 4*(1), 26–42.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech and Language 15*(4), 403–434.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research 34*(4), 365–399.

Heylighen, F. and J.-M. Dewaele (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science 7*(3), 293–340.

Hovy, E. H. (1988). *Generating Natural Language under Pragmatic Constraints*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Isard, A., C. Brockmann, and J. Oberlander (2005). Re-creating dialogues from a corpus. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation at Corpus Linguistics 2005 (CL-05)*, Birmingham, UK, pp. 7–12.

Isard, A., C. Brockmann, and J. Oberlander (2006). Individuality and alignment in generated dialogues. In *Proceedings of the 4th International Natural Language Generation Conference (INLG-06)*, Sydney, Australia, pp. 22–29.

Isbister, K. and C. Nass (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human–Computer Studies 53*(2), 251–267.

Kuhn, R. and R. De Mori (1990). A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*(6), 570–583.

Langkilde, I. and K. Knight (1998). Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-98)*, Montréal, QB, Canada, pp. 704–710.

Mairesse, F. and M. Walker (2006a). Automatic recognition of personality in conversation. In *Proceedings of the 3rd Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-06), Short Papers*, New York, NY, USA, pp. 85–88.

Mairesse, F. and M. Walker (2006b). Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci2006)*, Vancouver, BC, Canada, pp. 543–548.

Mairesse, F. and M. Walker (2007). PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, Czech Republic, pp. 496–503.

Mairesse, F. and M. Walker (2008). Trainable generation of Big-Five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, Columbus, OH, USA, pp. 165–173.

Mairesse, F., M. A. Walker, M. R. Mehl, and R. K. Moore (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research 30*, 457–500.

Martin, D. L., A. J. Cheyer, and D. B. Moran (1999). The Open Agent Architecture: A framework for building distributed software systems. *Applied Artificial Intelligence 13*(1/2), 91–128.

Matthews, G., I. J. Deary, and M. C. Whiteman (2003). *Personality Traits* (Second ed.). Cambridge, UK: Cambridge University Press.

Mehl, M. R., S. D. Gosling, and J. W. Pennebaker (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology 90*(5), 862–877.

Minnen, G., J. Carroll, and D. Pearce (2001). Applied morphological processing of English. *Natural Language Engineering 7*(3), 207–223.

Moore, J., M. E. Foster, O. Lemon, and M. White (2004). Generating tailored, comparative descriptions in spoken dialogue. In *Proceedings of the 17th International FLAIRS Conference (FLAIRS-17)*, Miami Beach, FL, USA.

Nass, C. and K. M. Lee (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied 7*(3), 171–181.

Niederhoffer, K. G. and J. W. Pennebaker (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology 21*(4), 337–360.

Nowson, S. (2006). *The Language of Weblogs: A study of genre and individual differences*. Ph. D. thesis, University of Edinburgh, UK.

Oberlander, J. and A. J. Gill (2004). Individual differences and implicit language: personality, parts-of-speech and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society (CogSci2004)*, Chicago, IL, USA, pp. 1035–1040.

Oberlander, J. and A. J. Gill (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes 42*(3), 239–270.

Oberlander, J. and S. Nowson (2006). Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-06) Main Conference Poster Sessions*, Sydney, Australia, pp. 627–634.

Paiva, D. S. and R. Evans (2004). A framework for stylistically controlled generation. In *Proceedings of the 3rd International Natural Language Generation Conference (INLG-04)*, Brockenhurst, UK, pp. 120–129.

Pearson, J., J. Hu, H. P. Branigan, M. J. Pickering, and C. I. Nass (2006). Adaptive language behavior in HCI: How expectations and beliefs about a system affect users' word choice. In *Proceedings of the 24th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI-06)*, Montréal, QB, Canada, pp. 1177–1180.

Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual*. Mahwah, NJ, USA: Erlbaum Publishers.

Pennebaker, J. W. and L. A. King (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology 77*(6), 1296–1312.

Pickering, M. J. and H. P. Branigan (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language 39*(4), 633–651.

Pickering, M. J. and V. S. Ferreira (2008). Structural priming: A critical review. *Psychological Bulletin 134*(3), 427–459.

Pickering, M. J. and S. Garrod (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences 27*(2), 169–225.

Pickering, M. J. and S. Garrod (2006). Alignment as the basis for successful communication. *Research on Language and Computation 4*(2–3), 203–228.

Piwek, P. (2003). A flexible pragmatics-driven language generator for animated agents. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03), Research Notes*, Budapest, Hungary, pp. 151–154.

Rammstedt, B. and J. Schupp (2008). Only the congruent survive – personality similarities in couples. *Personality and Individual Differences 45*(6), 533–535.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing (EMNLP-1)*, Philadelphia, PA, USA, pp. 133–142.

Reeves, B. and C. Nass (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Stanford, CA, USA: CSLI Publications.

Reiter, E. and R. Dale (2000). *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press.

Reitter, D. (2008). *Context Effects in Language Production: Models of Syntactic Priming in Dialogue Corpora*. Ph. D. thesis, University of Edinburgh, UK.

Shepard, C. A., H. Giles, and B. A. Le Poire (2001). Communication accommodation theory. In W. P. Robinson and H. Giles (Eds.), *The New Handbook of Language and Social Psychology*, Chapter 1.2, pp. 33–56. Chichester, UK: John Wiley & Sons.

Steedman, M. (2000). *The Syntactic Process*. Language, Speech, and Communication. Cambridge, MA, USA: MIT Press.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-02)*, Denver, CO, USA, pp. 901–904.

Stolcke, A., H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng (2000). The SRI March 2000 Hub-5 conversational speech transcription system. In *Proceedings of the 2000 Speech Transcription Workshop*, College Park, MD, USA.

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory 1*(1), 113–150.

van Deemter, K., B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann (2008). Fully generated scripted dialogue for embodied agents. *Artificial Intelligence 172*(10), 1219–1244.

Varges, S. and C. Mellish (2001). Instance-based natural language generation. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, Pittsburgh, PA, USA, pp. 1–8.

Walker, M., R. Prasad, and A. Stent (2003). A trainable generator for recommendations in multimodal dialog. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Interspeech-03 – Eurospeech)*, Geneva, Switzerland, pp. 1697–1700.

White, M. (2006a). CCG chart realization from disjunctive inputs. In *Proceedings of the 4th International Natural Language Generation Conference (INLG-06)*, Sydney, Australia, pp. 9–16.

White, M. (2006b). Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation 4*(1), 39–75.

Wilson, M. D. (1988). The MRC Psycholinguistic Database: Machine readable dictionary, version 2. *Behavior Research Methods, Instruments, & Computers 20*(1), 6–11.